Critical Infrastructure Protection in Homeland Security

Defending a Networked Nation

Third Edition

Ted G. Lewis

with website

Wiley

CRITICAL INFRASTRUCTURE PROTECTION IN HOMELAND SECURITY

CRITICAL INFRASTRUCTURE PROTECTION IN HOMELAND SECURITY

Defending a Networked Nation

Third Edition

TED G. LEWIS

WILEY

This third edition first published 2020

Edition History
John Wiley & Sons Inc. (1e, 2006)
John Wiley & Sons Inc. (2e, 2015)

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Lewis, T. G. (Theodore Gyle), 1941– author.
Title: Critical infrastructure protection in homeland security : defending a networked nation / Theodore Gyle Lewis.
Description: Third edition. | Hoboken, NJ : John Wiley & Sons Inc., 2020. | Includes bibliographical references and index.
Identifiers: LCCN 2019032791 (print) | LCCN 2019032792 (ebook) | ISBN 9781119614531 (hardback) | ISBN 9781119614555 (adobe pdf) | ISBN 9781119614562 (epub)
Subjects: LCSH: Computer networks–Security measures–United States. | Computer security–United States–Planning. | Terrorism–United States–Prevention. | Terrorism–Government policy–United States. | Civil defense–United States. | Public utilities–Protection–United States.
Classification: LCC QA76.9.A25 L5 2020 (print) | LCC QA76.9.A25 (ebook) | DDC 005.8–dc23
LC record available at https://lccn.loc.gov/2019032791
LC ebook record available at https://lccn.loc.gov/2019032792

Cover design by Wiley
Cover image: © SERGII IAREMENKO/SCIENCE PHOTO LIBRARY/Getty Images

Set in 10/12pt Times by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

FOREWORD BY SEN. MARK WARNER

"Today, December 7th, is an auspicious date in our history. We remember Pearl Harbor as the first foreign attack on US soil in modern history. Unfortunately, we also remember Pearl Harbor as a major intelligence failure. As Vice Chairman of the Intel Committee, I've spent the better part of the last two years on an investigation connected to America's most recent intelligence failure. It was also a failure of imagination—a failure to identify Russia's broader strategy to interfere in our elections. Our federal government and institutions were caught flat-footed in 2016, and our social media companies failed to anticipate how their platforms could be manipulated and misused by Russian operatives. Frankly, we should have seen it coming.

Over the last two decades, adversary nations like Russia have developed a radically different conception of information security—one that spans cyber warfare and information operations. I fear that we have entered a new era of nation-state conflict: one in which a nation projects strength less through traditional military hardware and more through cyber and information warfare. For the better part of two decades, this was a domain where we thought we had superiority. The thinking was that our cyber capabilities were unmatched. Our supposed superiority allowed us to write the rules.

This confidence appears to have blinded us to three important developments: First, we are under attack, and candidly, we have been for many years. Our adversaries and their proxies are carrying out cyber attacks at every level of our society. We've seen state-sponsored or sanctioned attacks on healthcare systems, energy infrastructure, and our financial system. We are witnessing constant intrusions into federal networks. We're seeing regular attempts to access parts of our critical infrastructure and hold them ransom. Last year, we saw global ransomware attacks increase by 93%. Denial-of-service attacks increased by 91%. According to some estimates, cyber attacks and cybercrime account for up to $175 billion in economic and intellectual property loss per year in North America. Globally, that number is nearly $600 billion. Typically, our adversaries aren't using highly sophisticated tools. They are attacking opportunistically using phishing techniques and rattling unlocked doors. This has all been happening under our noses. The effects have been devastating, yet the attackers have faced few, if any, consequences.

Second, in many ways, we brought this on ourselves. We live in a society that is becoming more and more dependent on products and networks that are under constant attack. Yet the level of security we accept in commercial technology products is unacceptably low—particularly when it comes to rapidly growing Internet of Things. This problem is only compounded by our society-wide failure to promote cyber hygiene. It is an outrage that more digital services from email to online banking don't come with default two-factor authentication. And it is totally unacceptable that large enterprises—including federal agencies—aren't using the available tools.

Lastly, we have failed to recognize that our adversaries are working with a totally different playbook. Countries like Russia are increasingly merging traditional cyber attacks with information operations. This emerging brand of hybrid cyber warfare exploits our greatest strengths—our openness and free flow of ideas. Unfortunately, we are just now waking up to it. Looking back, the signs should have been obvious. Twenty years ago, Sergei Lavrov, then serving as Russia's UN Ambassador, advanced a draft resolution dealing with cyber and prohibiting particularly dangerous forms of information weapons. We can debate the sincerity of Russia's draft resolution, but in hindsight, the premise of

this resolution is striking. Specifically, the Russians saw traditional cyber warfare and cyber espionage as interlinked with information operations. It's true that, as recently as 2016, Russia continued to use these two vectors—cyber and information operations—on separate tracks. But there is no doubt that Putin now sees the full potential of hybrid cyber operations. By contrast, the United States spent two decades treating information operations and traditional information security as distinct domains. Increasingly, we treated info operations as quaint and outmoded. Just a year after Lavrov introduced that resolution, the United States eliminated the United States Information Agency, relegating counterpropaganda and information operations to a lower tier of foreign policy. In the two decades that followed, the United States embraced the Internet revolution as inherently democratizing. We ignored the warning signs outside the bubble of Western democracies.

The naïveté of US policy makers extended not just to Russia, but to China as well. Recall when President Clinton warned China that attempts to police the Internet would be like nailing Jell-O to the wall. In fact, China has been wildly successful at harnessing the economic benefits of the Internet in the absence of political freedom. China's doctrine of cyber sovereignty is the idea that a state has the absolute right to control information within its border. This takes the form of censorship, disinformation, and social control. It also takes the form of traditional computer network exploitation. And China has developed a powerful cyber and information affairs bureaucracy with broad authority to enforce this doctrine. We see indications of the Chinese approach in their successful efforts to recruit Western companies to their information control efforts. Just look at Google's recent push to develop a censored version of its search engine for China. Today, China's cyber and censorship infrastructure is the envy of authoritarian regimes around the world. China is now exporting both its technology and its cyber-sovereignty doctrine to countries like Venezuela, Ethiopia, and Pakistan. With the export of these tools and ideas, and with countries like North Korea and Iran copying Russia's disinformation playbook, these challenges will only get worse. And yet as a country we remain complacent.

Despite a flurry of strategy documents from the White House and DoD, the federal government is still not sufficiently organized or resourced to tackle this hybrid threat. We have no White House cyber czar, nor cyber bureau or senior cyber coordinator at the State Department. And we still have insufficient capacity at State and DHS when it comes to cybersecurity and disinformation. Our Global Engagement Center at the State Department is not sufficiently equipped to counter propaganda from our adversaries. And the White House has still not clarified roles and responsibilities for cyber across the US government. While some in the private sector have begun to grapple with the challenge, many more remain resistant to the changes and regulations needed. And the American people—still not fully aware of the threat—have not internalized the lessons of the last few years. We have a long way to go on cyber hygiene and online media consumption habits. Let me be clear: Congress does not have its act together either. We have no cyber committee. Cyber crosses numerous committee jurisdictions frequently hindering our ability to get ahead of the problem.

It's even worse in the area of misinformation/disinformation. The dangers are only growing as new technologies such as Deepfakes audio and video manipulation that can literally put words into someone's mouth are commercialized. The truth is, we are becoming ever more dependent on software. But at the same time, we are treating cybersecurity, network resiliency, and data reliability as afterthoughts. And these vulnerabilities will only continue to grow as our so-called real economy becomes increasingly inseparable from the digital economy.

If we're going to turn this around, we need not just a whole-of-government approach; we need a whole-of-society cyber doctrine. So what would a US cyber doctrine look like? It's not enough to simply improve the security of our infrastructure, computer systems, and data. We must also deal with adversaries who are using American technologies to exploit our freedom and openness and attack our democracy.

Let me lay out five recommendations:

1 NEW RULES

First, we need to develop new rules and norms for the use of cyber and information operations. We also need to better enforce existing norms. And most importantly, we need to do this on an international scale. We need to develop shared strategies with our allies that will strengthen these norms. When possible, we need to get our adversaries to buy into these norms as well. The truth is, our adversaries continue to believe that there won't be any consequences for their actions. In the post-9/11 national security environment, we spent tremendous energy combating terrorism and rogue states. But frankly, we've allowed some of our near-peer adversaries to operate with relative impunity when they attack the United States in the digital domain. There have been some reports in the press about the United States supposedly punching back at second-tier adversaries on occasion. But we've largely avoided this with Russia and China out of a fear of escalation. If a cyber attack shuts down Moscow for 24 h with no power, that's a problem. If someone were to shut down New York for 24 h, that would be a global crisis. As a result, for Russia and China, it's pretty much been open season on the United States. That has to end.

We need to have a national conversation about the defensive and offensive tools we are willing to use to respond

to the ongoing threats we face. In short, we need to start holding our adversaries accountable. Failing to articulate a clear set of expectations about when and where we will respond to cyber attacks is not just bad policy, but it is downright dangerous. We are allowing other nations to write the playbook on cyber norms. Part of this is the result of US inaction: from the late 1990s into the early 2000s, the United States was a consistent dissenting voice in UN meetings where cyber norms were proposed. In part, this reflected our aversion to piecemeal approaches to cybersecurity. But it also reflected a view that we didn't want to be bound by lesser powers. In 2015, there was a major effort at the UN—including the United States—to agree to principles of state behavior in cyberspace. We saw some international consensus around protecting critical infrastructure and investigating and mitigating cybercrime. Unfortunately, those 2015 principles at the UN failed to address economic espionage. And even the 2015 US–China cyber espionage deal was insufficient. And in 2017, disagreements between the United States, China, and Russia at the UN led to a deadlock on the question of how international law should apply to cyber conflicts. Little progress has been made since then.

It's true that some folks in the private sector and the NGO space have stepped up. Look at Microsoft's Digital Geneva Convention. Look at the recent Paris Call for Trust and Security in Cyberspace—signed by 57 nations, but not by the United States. This is yet another example of the United States stepping back on the world stage, with countries like France filling the void.

Recently, the US government and the State Department, in particular, have renewed efforts to advance a norms discussion. These efforts must be elevated and strengthened. But norms on traditional cyber attacks alone are not enough. We also need to bring information operations into the debate.

This includes building support for rules that address the Internet's potential for censorship and repression. We need to present alternatives that explicitly embrace a free and open Internet. And we need that responsibility to extend not only to government, but to the private sector as well. We need multilateral agreements with key allies, just like we've done with international treaties on biological and chemical weapons. That discussion needs to address mutual defense commitments.

We should be linking consensus principles of state behavior in cyberspace, explicitly, with deterrence and enforcement policies. US policy makers, with allies, should predetermine responses for potential targets, perpetrators, and severity of attack. That means clearly and publicly linking actions and countermeasures to specific provocations. That could mean sanctions, export controls, or indictments. It could even include military action or other responses. Now, we should be realistic about the limits of norms in shaping behavior.

Let's not kid ourselves: in the short term, a nation like Russia that routinely ignores global norms is not going to make an about-face in the cyber domain. This should not deter us, but it should give us a more realistic set of expectations for how quickly we can expect to see results. But the stronger we make these alliances, the more teeth we can apply to these norms, and the more countries we can recruit to them, the more effective these efforts will be at disciplining the behavior of Russia, China, and other adversaries.

2 COMBATING MISINFORMATION AND DISINFORMATION

My second recommendation is: we need a society-wide effort to combat misinformation and disinformation, particularly on social media. My eyes were really opened to this through the Intel Committee's Russia investigation. Everyone on the Committee agrees that this linkage between cyber threats and disinformation is a serious challenge—especially on social media. In some ways, this was a whole new world for the IC. It is now clear that foreign agents used American-made social media to spread misinformation and hijack our civil discourse.

Let's recap. The Russian playbook included:

- Cyber penetrations of our election infrastructure;
- Hacks and weaponized leaks;
- Amplification of divisive, pro-Kremlin messages via social media;
- Overt propaganda;
- Funding and supporting extreme candidates or parties; and
- Misinformation, disinformation, and actual fake news.

The goal was, and is, to undermine our faith in the facts—our faith in the news media—and our faith in the democratic process. This is an ongoing threat, and not just to the United States. We've also seen these tools used against other Western democracies. We've seen them used to incite racial and ethnic violence in places like Myanmar. This threat is particularly serious in countries with low media literacy. In many ways, social media IS the Internet in some of these countries. So, what do we do? How do we combat this threat? We can start by recognizing that this is a truly global problem. A twenty-first-century cyber and misinformation doctrine should lean into our alliances with NATO countries and other allies who share our values.

Earlier this year, Senator Rubio and I brought together a group of 12 parliamentarians from our NATO allies at the Atlantic Council. We held a summit focused on combating Russian election interference. Ironically, this was the very same day that our President stood on stage and kowtowed to

Vladimir Putin in Helsinki. Meanwhile, we were working with our NATO allies to develop a road map for increased cooperation and information sharing to counter Russian cyber and misinformation/disinformation aggression. In many cases, these countries are further along in educating their populations about the threat of misinformation and disinformation.

Last month, I met with the Prime Minister of Finland. As he put it, the Finns have been dealing with Russian misinformation and disinformation for over a 100 years. Finland is one of the most resilient countries when it comes to countering this threat from its neighbor to the east. Why is that? Again, it is their whole-of-society approach. It relies on a free press that maintains trust through strong self-regulatory mechanisms and journalistic standards. It places limits on social media platforms. They also have a vibrant digital civics initiative.

Finland's approach also depends on national leadership that stays true to its values—even in the midst of contested elections and its own brand of partisan politics. Here in the United States, it will take all of us—the private sector, the government, including Congress, and the American people—to deal with this new and evolving threat.

In terms of the private sector, the major platform companies—like Twitter and Facebook, but also Reddit, YouTube, and Tumblr—aren't doing nearly enough to prevent their platforms from becoming petri dishes for Russian disinformation and propaganda.

I don't have any interest in regulating these companies into oblivion. But as these companies have grown from dorm-room startups into media behemoths, they have not acknowledged that their power comes with great responsibility. Recall that immediately following the election, Mr. Zuckerberg publicly ridiculed the idea that Russia had influenced the US election via Facebook as a "pretty crazy idea."

Now, I don't have all the solutions. But I expect these platforms to work with us in Congress so that together we can take steps to protect the integrity of our elections and our civil discourse in the future. Companies like Facebook and Twitter have taken some helpful voluntary steps—but we need to see much more from them.

That's going to require investments in people and technology to help identify misinformation before it spreads widely. I've put forward a white paper, which lays out a number of policy proposals for addressing this: we can start with greater transparency. For example, I think folks have the right to know if information they're receiving is coming from a human or a bot. I've also put forward legislation called the Honest Ads Act that would require greater transparency and disclosure for online political ads.

Companies should also have a duty to identify inauthentic accounts—if someone says they're Mark from Alexandria but it's actually Boris in St. Petersburg, I think people have a right to know. We also need to put in place some consequences for social media platforms that continue to propagate truly defamatory content. I think platforms should give greater access to academics and other independent analysts studying social trends like disinformation. We also discuss in that paper a number of other ideas in the white paper around privacy, price transparency, and data portability. These are ideas intended to spark a discussion, and we need social media companies' input. But we're moving quickly to the point where Congress will have no choice but to act on its own. One thing is clear: the wild west days of social media are coming to an end.

3 HARDEN NETWORKS, WEAPONS SYSTEMS, AND IOT (INTERNET OF THINGS)

Third, we need to harden the security of our computer networks, weapons systems, and IoT devices. Many of the responsibilities for cyber and misinformation/disinformation will fall on the government. But our nation's strategic response must also include greater vigilance by the private sector, which has frequently resisted efforts to improve the security of its products.

For over a decade, the United States thought it could set a light-touch standard for global data protection by avoiding any legislation. While regulation can have costs, what we've learned is that US inaction can also have costs—as other jurisdictions leap ahead with more stringent privacy and data protections.

We see this with GDPR, where the US failure to adopt reasonable data protection and privacy rules left the field open for much stricter European rules. These standards are now being adopted by major economies like Brazil, India, and Kenya. More broadly, we need to think about a software liability regime that drives the market toward more secure development across the entire product lifecycle. But nowhere is the need for private sector responsibility greater than the Internet of Things. General Ashley, Director of the DIA, has described insecure IoT and mobile devices as the most important emerging cyber threat to our national security.

As a first step, we should use the purchasing power of the federal government to require that devices meet minimum security standards. I have legislation with Senator Cory Gardner to do this. At least at the federal level, we need to make sure that these devices are patchable. We need to make sure they don't have hard-coded passwords that cannot be changed. We need standards to make sure they're free of known security vulnerabilities. And on a broader level, public companies should have at least one board member who can understand and model cyber risk.

Another area I've been working on is trying to impose some financial penalties on companies like Equifax who fail

to take the necessary steps to secure their systems from cyber intrusions. Unfortunately, even in areas where we would expect a higher level of security and cyber hygiene, we find these same problems. In October, a GAO report found that "nearly all" of our new weapons systems under development are vulnerable to attack.

Earlier this year, we successfully included language in the NDAA requiring cyber vulnerability assessments for weapons systems, which hopefully should help correct this. The Pentagon has also taken steps recently to make cybersecurity a greater priority within DoD, but frankly we face some serious workforce challenges in recruiting and retaining the top cyber professionals who have plenty of lucrative opportunities in the private sector.

4 REALIGN DEFENSE SPENDING

This is a good segue to my fourth recommendation: realigning our defense spending priorities. The US military budget is more than $700 billion, while Russia spends roughly $70 billion a year on their military. The United States is spending it mostly on conventional weapons and personnel. By contrast, Russia devotes a much greater proportion of its budget to cyber and other tools of asymmetric warfare like disinformation. Russia has come to the realization that they can't afford to keep up with us in terms of traditional defense spending. But when it comes to cyber, misinformation, and disinformation, candidly Russia is already a peer adversary.

A matter of fact, if you add up everything Russia spent on election interference in 2016 and double it, that's still less than the cost of one new F-35. I worry we may be buying the world's best twentieth-century military hardware without giving enough thought to the twenty-first-century threats we face. And it's a similar story with China. China spends roughly $200 billion on defense, but it spends a greater proportion on cyber misinformation and disinformation. If you look at the delta between what we're spending and what China is spending on defense, they're investing more in AI, quantum computing, 5G, and other twenty-first-century technologies. Frankly, they are outpacing us by orders of magnitude. We need to realign our priorities while we still can. Some of DoD's budget should be redirected toward cyber defense. But we also need efforts at other agencies, including R&D funding for quantum computing and AI, as well as investments in cyber technology and cyber workforce development.

5 PRESIDENTIAL/GOVERNMENT LEADERSHIP

The final point is that we desperately need strong federal and presidential leadership for any US cyber doctrine to be truly effective. Because this challenge literally touches every aspect of our society, we need presidential leadership and a senior coordinating official to head the interagency process on this issue.

It's true there are men and women within DoD, DHS, and other agencies who are working hard to defend the United States from cyber attacks. But only the President can mobilize the whole-of-society strategy we need. I do want to acknowledge some positive steps that have been taken in recent months.

The White House and DoD have released two important strategic documents on cyber strategy that move us in the right direction. I also welcome the delegation of authorities to defend and deter cyber attacks below the presidential level. This has allowed for quicker responses and greater interagency coordination. But frankly, these efforts are inadequate.

In the most recent NDAA, Congress attempted to establish a more aggressive posture on US cybersecurity policy. This includes the potential use of offensive cyber capabilities to deter and respond to cyber attacks against US interests—as well as authorization to combat info operations. It also grants the President and Defense Secretary authority to direct Cyber Command to respond and deter "an active, systematic, and ongoing campaign of attacks" carried out by Russia, China, North Korea, and Iran. These powers, if used correctly, are important components of a cyber doctrine. But by definition they require thoughtful, decisive leadership at the top.

I'll leave you with some final thoughts. More broadly, we need a coherent strategy for how to deal with the hybrid approach of our adversaries. Let me be clear about what I'm not saying: I am not advocating that the United States mimic the approach of Russia and China—the idea that states have a sovereign right to control or censor information within their borders. Frankly, that vision is incompatible with our American values and our Constitution.

What I am saying is that we need to confront the fact that our adversaries have an approach that considers control of information an essential component of their overall strategies. We have not only failed to recognize this situation, but over the last two decades we have tended to minimize the dangers of information operations. The truth is, the 2016 presidential election served as a wake-up call in the use of cyber attacks and information operations.

People keep warning of a "digital Pearl Harbor" or a "digital 9/11" as if there will be a single extraordinary event that will force us to action on these issues. But I have news for you: we are already living these events. They're happening every day. Look at the 2017 NotPetya attack. In the United States, we treated this as a one-day news story, but the global cost of that one attack is over $10 billion. This is the most costly and devastating cybersecurity incident in history, and most Americans have no idea. But the true costs of

our cyber vulnerabilities won't be sudden or catastrophic. They will be gradual and accumulating. Our personal, corporate, and government data is being bled from our networks every day; our faith in institutions and our tolerance for one another is being eroded by misinformation. This is leaving us exposed as individuals and vulnerable as a country. It's time we dramatically shift how we view these threats. I hope the ideas I've laid out today will help us move toward the comprehensive cyber doctrine that we so desperately need in these challenging times."

FOREWORD BY PROF. ANDREW ODLYZKO

Cybersecurity Is Not Very Important
Andrew Odlyzko
University of Minnesota
odlyzko@umn.edu http://www.dtc.umn.edu/∼odlyzko Revised version, March 9, 2019.

A New Doctrine for Cyberwarfare and Information Operations, Center for New American Security, Sen. Mark R. Warner, December 7, 2018

1 INTRODUCTION

It is time to acknowledge the wisdom of the "bean counters." For ages, multitudes of observers, including this author, have been complaining about those disdained accountants and business managers. They have been blamed for placing excessive emphasis on short-term budget constraints, treating cybersecurity as unimportant, and downplaying the risks of disaster.

With the benefit of what is now several decades of experience, we have to admit those bean counters have been right. The problems have simply not been all that serious. Further, if we step back and take a sober look, it becomes clear those problems are still not all that serious.

All along, the constant refrain has been that we need to take security seriously and engineer our systems from the ground up to be truly secure. The recent report [3] opens with a quote from a 1970 publication (the well-known Ware Report) that called for such moves. This demand has been growing in stridency and has been increasingly echoed by higher levels of management and of political leadership. Yet in practice over the last few decades, we have seen just a gradual increase in resources devoted to cybersecurity. Action has been dominated by minor patches. No fundamental reengineering has taken place.

This essay argues that this "muddle-through" approach was not as foolish as is usually claimed and will continue to be the way we operate. Cyber infrastructure is becoming more important. Hence intensifying efforts to keep it sufficiently secure to let the world function is justified. But this process can continue to be gradual. There is no need to panic or make drastic changes, as the threats are manageable and not much different from those that we cope with in the physical realm.

This essay reviews from a very high level the main factors that have allowed the world to thrive in spite of the clear lack of solid cybersecurity. The main conclusion is that through incremental steps, we have in effect learned to adopt techniques from the physical world to compensate for the deficiencies of cyberspace. This conclusion is diametrically opposed to the heated rhetoric we observe in the popular media and to the unanimous opinions of the technical and professional literature. No claim is made that this process was optimal—just that it was "good enough." Further, if we consider the threats we face, we are likely to be able to continue operating in this way. But if we look at the situation realistically, and plan accordingly, we might:

- Enjoy greater peace of mind
- Produce better resource allocations

The analysis of this essay does lead to numerous contrarian ideas. In particular, many features of modern technologies such as "spaghetti code" or "security through obscurity" are almost universally denigrated, as they are substantial contributors to cyber insecurity. But while this is true, they are

also important contributors to the imperfect but adequate levels of cybersecurity that we depend on. Although a widely cited mantra is that "complexity is the enemy of security," just the opposite is true in the world we live in, where perfect security is impossible. Complexity is an essential element of the (imperfect) security we enjoy, as will be explained in more detail later. Hence one way to improve our security is to emphasize "spaghetti code" and "security through obscurity" explicitly and implement them in systematic and purposeful ways. In general, we should adopt the Dr. Strangelove approach, which is to stop worrying and learn to love the bomb.

In other words, not just accept that our systems will be insecure. Recognize that insecurity often arises in systematic ways and that some of those ways can be turned into defensive mechanisms. We do have many incremental ways to compensate, and we have to learn how to systematically deploy them, so as to live and prosper anyway. The key point is that, in cyberspace as well as in physical space, security is not the paramount goal by itself. Some degree of security is needed, but it is just a tool for achieving other social and economic goals.

Historically, for many observers, a serious reassessment of the traditional search for absolute security was provoked by Dan Geer's 1998 post [1]. However, awareness of general risk issues, and growing perception that they were key, can be traced much further back to various research efforts in the 1980s and the founding of Peter Neumann's RISKS Digest in 1985. No attempt is made here to trace this evolution of attitudes toward security. That is a nice large subject that is left for future historians to deal with. This essay considers only the current situation and likely evolution in the near future.

2 THE TECHNOLOGISTS' SKEWED VIEW OF THE WORLD

The critics of the standard "business as usual" approach have been presenting to the public both a promise and a threat. The promise was that with enough resources and control over system development, truly secure information technologies systems would be built. The threat was that a gigantic disaster, a "digital Pearl Harbor," would occur otherwise.

The promise of real security was hollow. If there is anything that we can now regard as solidly established, it is that we don't know how to build secure systems of any real complexity. (There is another factor that is not discussed here, namely, that even if we could build truly secure systems, we probably could not live with them, as they would not accommodate the human desires for flexibility and ability to bend the rules. But that is a different issue not in the scope of this essay.) Serious bugs that pose major security risks are being found even in open-source software that has been around and in extensive use for years, as with the Heartbleed defect. And some insecurities, such as those revealed in the recent Meltdown and Spectre attacks, not only go back decades, but are deeply embedded in the basic architecture of modern digital processors. They cannot be eliminated easily, and we will have to live with them for many years. The most we can hope for is to mitigate their deleterious effects.

The mantra, called Linus's law, that "given enough eyeballs, all bugs are shallow" has been convincingly shown to be fallacious. There are only relative degrees of security. Still, we have to remember that this has always been true with physical systems. Furthermore, in both the cyber and the physical realms, the main vulnerabilities reside in people. Those creatures are not amenable to reengineering and are only very slightly amenable to reasoning and education.

The threat of digital catastrophe has also turned out to be hollow. Sherlock Holmes noted that the "curious incident" in the Silver Blaze story was that the dog did not bark. In information technology insecurity, there are two curious "incidents" that have not attracted much notice:

- Why have there been no giant cybersecurity disasters?
- Why is the world in general doing as well as it is?

Skeptics might object and point out to any number of ransomware, identity theft, and other cybercrime cases. But those have to be kept in perspective, as is argued in more detail later. There have been many far larger disasters of the non-cyber kind, such as 9/11, Hurricane Sandy, the Fukushima nuclear reactor meltdown, and the 2008 financial crash and ensuing Great Recession. Has any cyber disaster inflicted anywhere near as much damage to any large population as Hurricane Maria did to Puerto Rico in 2017?

In the cyber realm itself, we have experienced many prominent disasters. But most of them, such as airlines being grounded for hours or days or cash machine networks not functioning, have arisen not from hostile action, but from ordinary run-of-the-mill programming bugs or human operational mistakes. And of course we have the myriad issues such as cost overruns and performance disappointments which plague information as well as other rapidly evolving technologies. They have little to do with the lack of cybersecurity. Yet we suffer from them every day.

There is a third curious incident in information technology (in)security that also appears to be universally ignored. For several decades we have had simple tools for strengthening security that did not require any fundamental reengineering of information systems. A very conspicuous example of such tools is two-factor authentication. The widely cited and widely accepted explanation for this technology not having been deployed more widely before is that users disliked the extra bother it involved. So apparently decision makers felt that the extra security provided by

two-factor authentication did not warrant the cost of inconveniencing users. The big "dog did not bark" question then is, given that this technology was not deployed, why did nothing terrible happen?

The general conclusion of this essay is that from the start, the "bean counters" understood the basic issues better than the technologists, even though they usually did not articulate this well. The main problem all along was risk mitigation for the human world in which cyberspace played a relatively small role; it was not absolute security for the visionary cyberspace that technologists dreamed of.

3 THE STATE OF THE WORLD

One could object that the world is not doing well and point to climate change, rising inequality, civil wars, unemployment, and other phenomena that are cited as major ills of our society. But that has to be kept in perspective. Let's put aside, until the next section, questions about issues such as long-term sustainability of our civilization. If we just look at where the human race is today from a long-term historical perspective, we find stunning advances by many measures, such as the number of people on Earth, how long they live, and how educated they are. There are more people today who are obese than hungry, which is unprecedented. Obesity is certainly not ideal, but can easily be argued to be an advance on the historically dominant feature of human lives.

Of course, there are a variety of threats for the future. But we need to remember that the progress that has occurred has relied often and in crucial ways on information systems that were, and are, insecure. Further, almost all of the most serious threats, to be considered next, are little affected by cybersecurity or lack of it.

4 THREATS

We certainly do face many threats. In particular, we do face many cyber threats. It seems inevitable that we will suffer a "digital Pearl Harbor." What we have to keep in mind is that we have suffered a physical Pearl Harbor and other non-cyber disasters that large or larger. Many occurred quite recently, as noted before. It seems absolutely certain we will suffer many more, and an increasing number of them will surely be coming from the cyber realm. On the other hand, it is questionable whether the cyber threats are yet the most urgent ones.

The human race faces many potentially devastating non-cyber dangers, such as asteroid strikes, runaway global warming, and large pandemics. These threats could have giant impacts, but are hard to predict and quantify and are seemingly remote, so tend to be ignored by almost all people most of the time. However, we also face a variety of other still large dangers, such as those from earthquakes and hurricanes. Those occur more frequently, so the damage they cause is moderately predictable, at least in a long-run statistical sense. Yet we are not doing anywhere near as much to protect against them as we could, if we wanted to do so. We accept that they will occur and rely on general resilience and insurance, whether of the standard variety, or the implicit insurance of governments stepping in with rescue and recovery assistance.

We also tolerate the ongoing slaughter of over a million people each year in automobile accidents worldwide (with about 40,000 in the United States alone). The horrendous losses of human life as well as property that involve cars arise mostly from unintentional mistakes. They result from our accepting the limitations of *Homo sapiens* when dealing with a dangerous technology. It's just that this technology has proven extremely attractive to our species. Hence we accept the collateral damage that results from its use, even though it far exceeds that from all wars and civil conflicts of recent times.

On top of accidents we also have the constant ongoing malicious damage, coming from crime in its many dimensions. Society suffers large losses all the time, and mitigates the threat, but has never been able to eliminate it. We have large security forces, criminal courts, jails, and so on. The United States alone has close to a million uniformed police officers and more than a million private security guards.

Military establishments tend to be substantially larger than law enforcement ones. The main justification for them is to guard against the far rarer but potentially more damaging actions of hostile nations. One way or another, most societies have decided to prioritize protection against those external dangers over that of internal crime. Further, in recent decades, military spending (and therefore total security-related spending) has been declining as a fraction of the world's economic output. So when societies feel threatened enough, they do manage to put far more effort into security than is the case today.

Yet even military security at its very best is not watertight, which has to be kept in mind when considering cybersecurity. Serious gaps have been uncovered on numerous occasions, such as a deep penetration of an American nuclear weapons facility by a pacifist group that included an 82-year-old nun.

The bottom line is that society has always been devoting huge resources to security without ever achieving complete security. But those huge resources are still not as great as they could be. That's because, as noted above, security is not the paramount goal by itself. We make trade-offs and are only willing to give up a fraction of the goods and services we produce for greater safety. There is even extensive evidence for human desire for a certain level of risk in their lives. When some safety measures are introduced, people compensate for that by behaving with less care.

Still, we do employ many people and extensive resources protecting ourselves from traditional physical world threats, far more than we devote to cybersecurity. Hence it is clear, and has been clear for a long time, that more effort could have been dedicated to cybersecurity, even without consuming productive resources. All we had to do was just shift some of the effort devoted to traditional physical security to the cyber realm. And indeed that is what is happening now, at least in relative sense. More attention and resources is being devoted to cybersecurity. One measure of the greater stress being placed on this area is the growing (but still very small) number of CEOs who have lost their jobs as result of security breaches. So the question arises, essentially the same question as before, just in a different form: Why was this not done before, and why has not much harm come from this?

5 HUMANSPACE VERSUS CYBERSPACE

It is very hard for technologists to give up the idea of absolute cybersecurity. Their mind-set is naturally attracted to the binary secure/insecure classification. They are also used to the idea of security being fragile. They are not used to thinking that even a sieve can hold water to an extent adequate for many purposes. The dominant mantra is that "a chain is only as strong as its weakest link." Yet that is probably not the appropriate metaphor. It is better to think of a net. Although it has many holes, it can often still perform adequately for either catching fish or limiting inflow of birds or insects. A tight sieve can even retain a substantial amount of water for a while.

Technologists also tend to think of information systems as isolated. This attitude is represented beautifully by the famous 1996 creation of John Perry Barlow: "A Declaration of the Independence of Cyberspace." This proclamation, which today seems outlandishly ludicrous, proclaimed the existence of a new realm, "cyberspace," that was divorced from the physical world and did not need or want traditional governments or other institutions. The key assumption was nicely formulated in the oft-quoted passage:

Cyberspace consists of transactions, relationships, and thought itself, arrayed like a standing wave in the web of our communications. Ours is a world that is both everywhere and nowhere, but it is not where bodies live.

Indeed, if cyberspace were totally divorced from humanspace, and if all the "transactions, relationships, and thought itself" depended just on some mathematical relationships, then cybersecurity would be of paramount importance. An opponent utilizing a clever mathematical idea to break a public key system, or stealing a password, might wreak unlimited havoc.

And indeed, as the increasing number of incidents with bitcoin and other cryptocurrencies proves, such dangers do lurk in pure cyber realms. Further, they cannot be avoided. As was discussed before, people are incapable of building completely secure systems, they do choose weak passwords or leak strong ones, they do fall prey to phishing attacks, and every once in a while a mathematical breakthrough does demolish a cryptosystem.

What makes our lives tolerable is that the Barlow vision is divorced from reality. Cyberspace is intimately tied to what we might call humanspace, the convoluted world of physical objects and multiple relations, including institutions such as governments, and laws, and lawyers. In fact, we can say:

> The dream of people like Barlow was to build a cyberspace that would overcome the perceived defects of humanspace. In practice we have used the defensive mechanisms of humanspace to compensate for the defects of cyberspace.

Those defensive mechanisms are what we consider next, starting with the limitations of attackers in both physical and cyber realms.

6 PLUSES AND MINUSES OF NATURAL STUPIDITY

There are extensive discussions going on about the promises and threats of artificial intelligence (AI). Much less is said about natural stupidity and its positive aspects. Yet it is central to human life and key to enabling society to function. (At an even more basic level, the astounding level of human credulity, which enables so many attacks, is an essential element of human psychology and sociology and enables the cooperation that has led to modern civilization.) In particular, we are alive and living pretty well largely because most criminals are stupid.

This includes terrorists. Most of them are stupid, too. They are in almost all cases more like the Shoe Bomber than the highly trained and highly proficient professionals that the multitudes of publicly prominent cyber Cassandras hold out as big threats to our lives. Most crimes are extremely mundane, and many more could easily be solved if more effort was devoted to them. Criminals constantly make foolish mistakes, such as leaving their fingerprints, or their DNA, on the scene or driving their own cars. As a result, general crime has been kept within tolerable bounds for most of human history.

It is not just the most stupid people who make mistakes. Everyone does so. In fact, the mistakes of the smartest individuals are often the most disastrous, as they get entrusted with the most important jobs. Even the highly trained and highly proficient professionals in the military and intelligence agencies are fallible, including when at the peak of training and preparation. It is this fallibility that helps make cyberspace more similar to physical space than

is commonly thought. Detecting where a network attack originates is harder than detecting where a ballistic missile is launched from. But digital forensics is a thriving field, largely because of human mistakes. Even the Stuxnet creators were not able to completely erase their "digital fingerprints," leading to high confidence as to their identities.

Cybercrimes not only leave digital fingerprints. They are usually tied in one way or another to the physical world, most frequently through flows of money. Hence there are far more ways to trace them than would be the case if they happened purely in cyberspace. Once tracing is possible, measures to deter, prevent, and punish can be brought to bear. Those digital fingerprints also mean that natural stupidity of attackers has more opportunities to display itself. And that offers opportunities for defense and countermeasures, just as in the traditional environment.

7 SMART AND STUPID CRIMINALS

The reasons most criminals are stupid are worth considering. An important one is that we mostly hear of the criminals who get caught and that is not a perfectly representative sample. The smart ones avoid detection and capture. But the really smart ones mostly figure out it is far safer and more comfortable to stay close to the line of legality. Serious damage to the system as a whole, or even to many individual players, tends to provoke strong countermeasures. Some criminals even learn to be symbiotes and contribute positively to society.

An insightful analogy can be drawn with biology. A virus that kills the host instantly tends to perish, as it has little chance to spread. The more successful viruses (more successful in terms of being widespread) are like those for the common cold, which cause relatively small annoyances that serve primarily to help them propagate. Many parasites evolve to become symbiotes, and the study of commensal relationships is a thriving field with a variety of examples.

8 THE CYBERCRIME ECOSYSTEM

Most criminals, even among those on the extreme edge of the stupidity spectrum, have no interest in destroying the system they are abusing. They just want to exploit it to extract value for themselves out of it.

An amusing and instructive example of illicit cyber behavior that maintains the functioning of the system is provided by the ransomware criminals. Studies have documented the high level of "customer care" they typically provide. They tend to give expert assistance to victims who do pay up and have difficulty restoring their computers to the original state. After all, those criminals do want to establish "reputations" that will induce future victims to believe that payment of the demanded ransom will give them back control of their system and enable them to go on with their lives and jobs.

An extreme example of exploitation of cyber insecurity without causing noticeable damage is that of national intelligence agencies. They carry out extensive penetrations of a variety of government and commercial systems, but are usually just after limited pieces of information and try (and usually succeed) in staying inconspicuous. In most cases they exploit only a tiny fraction of what they acquire, precisely in order not to raise suspicions about their activities. Of course, their activities do involve other dangers, when they acquire control of systems for future large-scale hostile activities. But such penetrations by state actors have to be handled at state levels, similarly to what occurs in the physical realm.

There are certainly some malicious actors who simply want to inflict damage, whether it is against a person against whom they have a grudge or, especially in case of terrorists, against society at large. But even such people are generally not as dangerous in cyberspace as they could be. First of all, there are not that many of them. Second, they generally have limited skills and resources, and are mostly very foolish, and engage in foolish activities. The more rational among them choose their targets and methods for maximal effectiveness in achieving whatever nefarious purposes they have in mind. For terrorists, say, cyberspace is generally not very attractive as a target. Blocking people from withdrawing money from cash machines or even causing a blackout in a city does not carry as strong a message as blowing up airplanes, bringing down buildings, or causing blood to flow among spectators in a sports arena.

There is much concern about ongoing technology developments making the lack of cybersecurity far more dangerous, especially as more devices go online and IoT (the Internet of Things) becomes more pervasive. Those are valid concerns, but let us keep in mind that those ongoing technology developments are also creating or magnifying many physical dangers even without taking advantage of cyber insecurity. Just think of drones (or possibly imaginary drone sightings) shutting down airports recently or drones or self-driving cars delivering bombs in the future.

In general, and reinforcing earlier discussions, society has always faced manifold dangers from its members misusing various technologies. Deterrence, detection, or punishment, in addition to general social norms, is what has enable civilized human life to exist. Contrary to the cyberlibertarian visions of people like Barlow (or many modern advocates of bitcoin and blockchain), they are likely to be just as crucial in the future, if not more so.

Of course, as the old saying goes, bank robbers went after banks because that is where the money was. But now the money is in cyberspace. So that is where criminals are moving. And that is also where security resources are being redirected, completely natural and expected, and happening at a measured pace.

9 BLACK SWANS VERSUS LONG TAILS

Cybersecurity efforts are dominated by very mundane work, monitoring the automated probes of the network or attacks of the "script kiddies." And perhaps most prominent and most boring, but absolutely critical, is assisting legitimate users who have forgotten their passwords, which is exactly analogous to the state of traditional physical security. Much of the time of firefighters and police officers is devoted to rescuing kittens stuck high up trees or handling temporarily inebriated but otherwise perfectly respectable citizens.

The evolution of the cybersecurity field over the last few decades has led to wide recognition among its practitioners that threats cannot be entirely eliminated. There are frequent references to minimizing "the attack surface," for example. This reflects the reality that one can limit attacks and the damage they can do, but not get rid of them. More resources can be used to lessen threats. But those resources are costly, either in terms of the pay and equipment of the security professionals, or, what is typically much more important, in terms of constraints on the legitimate users. So one is led to look at optimizing the allocation of resources and studying and modifying the incentives. One outgrowth of such thinking on the academic side has been the rise of the field of economics of information security. It has produced a flourishing literature and a series of annual workshops. Together with all other academic and industry efforts, it fits into the basic philosophy that animates modern economics, namely, of studying systems in equilibrium. There is ongoing hostile activity that is counteracted by security measures, and the task is to select the optimal combination of those measures that fit within some budget constraints.

One could view such approaches as concentration on the "long tail" of security threats. There are many of them—they require large resources in the aggregate to deal with, but individually they pose limited and reasonably well understood dangers. Overall, their potential impact can be estimated and constrained by standard approaches.

But then, at the other end of the spectrum, there are the "black swans," the giant security breaches that cause major damage. Those don't fit into the equilibrium framework (just as catastrophic financial collapses don't fit into the standard economic equilibrium framework and have been almost entirely ignored by mainstream economists). But neither do the giant physical disasters, such as Pearl Harbor or Hurricane Katrina. Their damaging effects basically can only be mitigated by designing in general resilience.

Measures that provide resilience against cyber attacks are often the same as those against traditional physical attacks or against natural disasters. As just one example, there is much concern about the damage to the electric power grid that might be caused by malicious actors. But the worst scenarios along those lines are similar to what we are sure to suffer when something like the Carrington Event occurs. This was the giant geomagnetic solar storm that hit the Earth in 1859. It caused widespread failures of the telegraphs, the only electrical grids in existence at that time. Estimates are that if it were to recur today, it would cause damages in the trillions of dollars. And it is bound to recur some day!

The conclusion that emerges is again that cyberspace is not all that different from the more traditional physical space we are more used to. And security measures for the two are again similar.

10 NEGLECT OF OBVIOUS SECURITY MEASURES

The main thesis of this note—that cybersecurity is not very important—is illustrated nicely by the phenomenon of two-factor authentication. This technique is spreading. It is not a panacea, but there is general agreement that it offers significant enhancement to security.

But why is it only now that two-factor authentication is coming into widespread use? The basic technique is ancient by the standards of the information technology industry. Two and a half decades ago, it was used at my employer of that time. The hardware tokens came from one of several suppliers that were already in that line of business.

Yet even at my former employer, two-factor authentication was abandoned after a while, and in most places, it was never put into service in that era. So what has changed to finally make this technology used more widely? As often happens, it was likely a combination of factors:

- Threats have increased.
- Implementing two-factor authentication has become easier.

The old hardware tokens of the 1990s were not very expensive, but they had to be carried around (as opposed to receiving a text on a mobile phone that people have with them almost all the time, say), and they required typing in strings of arbitrary symbols. Now we can use short texts, or hardware tokens that plug into a computer, or else mobile phones that communicate with a nearby computer wirelessly. So while the monetary costs of the basic system have not changed dramatically, the costs to users have declined significantly. And, of course, the threats have increased, as noted above, so the incentives to use two-factor authentication have grown.

Yet even now, two-factor authentication is nowhere near universal. Further, most deployments of it at this time appear to use the least secure version of it, with texts to mobile phones. Practical attacks on this version have been developed and applied. The more secure versions with hardware tokens are used much less frequently. Obviously what is happening is that choices are being made, the additional

inconvenience to users being weighed against the likely losses from hostile penetrations. Even without any new technology breakthroughs, more secure versions of two-factor authentication can be deployed when they are seen as necessary. But they are clearly not being seen as necessary at present.

There are many more examples of relatively easy steps that have been available for a long time and can strengthen security without any fundamental reengineering of information systems or rearranging how society functions. Consider the adoption of chip credit cards. They have been universal in much of the world for years, but are only now taking over in the United States. The costs have been understood by the banking industry, and it was decided, through a messy process by various stakeholders, that they were too high until the perceived threats increased.

Electronic voting is another prominent example where simple and well-known steps would have provided greater security a long time ago. Experts have been arguing from the start that purely electronic voting basically cannot be made secure, at least not with feasible technology and the financial resources that are available or are likely to be made available. All the evidence that has been gathered over the years supports this view. Further, all the advantages of electronic voting (convenience, accessibility for those with handicaps, quick collection of results, etc.) can be obtained very easily, together with a much higher degree of security, through the use of printed records that are preserved in physical form. The additional costs that are involved are very modest and seem well worth it to most people who have examined the situation, including this author. Yet in many jurisdictions this simple solution is being ignored. And it has to be admitted that so far no serious abuses have been documented. What is likely to happen is that if some big scandal surfaces that is based on a cyber breach, political leaders will swing into action and find the resources to provide the obvious solution. (We should remember that big voting scandals do occur all the time, based on other aspects of the voting system, and they lead to responses that vary with circumstances.) But, as seems typical in human affairs, it will likely take a big scandal to cause this to happen.

Electronic voting provides an interesting illustration of a cyber insecurity that is not difficult to fix, but is not being fixed. It also provides an example of a common phenomenon, namely, that the fix involves stepping back to the traditional physical world, in this case of messy paper ballots. (The same could be said of chip cards.) In other words, the insecurity of the cyber realm is compensated by a measure from the brick-and-mortar world.

An even better example of reliance on physical world to compensate for defects in cybersecurity is that of passwords. They have been pronounced obsolete and dead many times, but are still ubiquitous. A key element in making them more tolerable in spite of their well-known weaknesses is the use of paper for users to write them down (or, preferably, to write down hints for those passwords or passphrases). The security field has finally been forced to admit that asking users to remember scores of complicated passwords (and change them every few months) is not going to work, not with the bulk of human users. But paper slips work out quite well, as physical wallets and purses do not get stolen all that often.

Notice that there are many other direct physical methods for increasing security. Air-gapped systems, isolated from the Internet, have been standard in high-security environments. They are again not absolutely secure, as the Stuxnet case demonstrates. But they do provide very high levels of security, as breaching them requires special skills and extensive effort (as the Stuxnet case demonstrates, again). At a simpler level, allowing certain operations (such as resetting the options on a router or another device) only through the press of a physical button on the device also limits what attackers can do.

Frequent backups serve to mitigate ransomware and many other attacks. They can be automated so that they do not impose any significant mental transaction costs on the users. They increase the reversibility of actions, which is a key component to security (but seems not to be understood by the advocates of bitcoin and other cryptocurrencies). And they are not expensive in terms of hardware. Of course, backups increase security only if they are not subverted. But there are a variety of ways to make backups more trustworthy, such as using write-only media (such as some optical disks) or special controllers that limit what operations can be done.

We should also remember there is one piece of advice that applies in both cyberspace and physical space: if it's dangerous, don't use it! Some very cautious organizations disable USB ports on their computers, but such organizations are rare. Email attachments are a notorious carrier for all sorts of malicious software. They could be blocked, but seldom are. All these examples show how society has in effect accepted obvious risks in order to get benefits of insecure information technology solutions.

11 SURVEILLANCE CAPITALISM AND LOSS OF PRIVACY

The analogy between cyber and physical security is strong, but there are certainly substantial differences. The one that appears to be cited most frequently is privacy. There was no absolute privacy in the past. In particular, there was always the most intractable problem of all, namely, that of insider disclosure. (According to an old saying, "two people can keep a secret, as long as one of them is dead.") But modern threats to privacy are orders of magnitude larger than those faced in the past. Further, as we move forward, our central

and giant problem is that potential leakers are proliferating at a rapid pace. Individuals can convey far more information now than in the past, as the Manning, Martin, and Snowden information torrents from NSA demonstrate. For the majority of people, though, the main threat comes in the shape of the many devices we use, which is increasing in numbers and in their capability to transmit information about us to others. The cell phone is the premier example, but increasingly so is our fitness tracker, our TV set, and our electric meter. Practically nothing that we will be doing can be assumed to be secret in the future. This will even apply to our physiological reactions, even ones we do not express, or may not consciously be aware of, since they might be discerned by various sensors.

Already today, the old mantra that "on the Internet, nobody knows you are a dog" has in practice been turned on its head. Many organizations know not only that you are a dog but also what breed of dog you are and what kind of fleas you have.

For the purposes of this essay, the key counterpoint to this line of argument is that this erosion of privacy we experience has little to do with cyber insecurity. Some of that erosion does come from illicit hacking of our systems, which is indeed facilitated by the insecurity of our information systems. But most of it comes by design, as providers of services and devices purposely build them to collect data about users for exploitation by those providers and their (almost universally concealed) networks of partners. (Even the illicit hacking of those devices, databases, and so on can occur only because of this huge and legal, even though usually obfuscated, data gathering.) Hence there are no improvements in cybersecurity that would by themselves make a measurable difference to the erosion of privacy that we experience. To the extent that society wants to preserve some semblance of privacy, other methods will have to be used, which likely will have to be based on laws and regulations and to some extent on technologies for users to protect themselves.

On the other hand, the erosion of privacy is a key element to maintaining tolerable levels of security in general. Tens or sometimes hundreds of millions of credit cards are routinely captured by criminals by compromises of databases. Yet the overall damages are limited and often dominated by the cost of arranging for replacement cards. The prices of stolen credit card credentials on the black market are low, on the order of a dollar or so each. The reason is that banks have developed techniques for detecting credit card fraud. Those are based on knowledge of users' patterns of behavior. A typical card holder is not an anonymous "standing wave" of Barlow's imagination, or some account even more anonymous than those involved in the not-all-that anonymous bitcoin operations. Instead, such a person is in most case an individual who mostly follows a staid routine in life and in commercial transactions, say, stopping by a particular coffee shop on the way to work or dropping in at a grocery store on the way back from work.

There are many measures that erode privacy, such as cross-device tracking (in which users are identified even though they use different gadgets) or identifying users by the patterns of their typing, that are often regarded as objectionable or even creepy. Yet they do serve to identify users, and thereby to prevent mischief, even if this is incidental to the main purposes for which they are deployed. Organizations that operate these systems can get a high degree of assurance as to the person they are dealing with and in such circumstances stealing a credit card or cracking a password is often of limited use.

It should also be remembered that since enterprises do want to track customers or potential customers for their own business reasons, they have incentives to develop and deploy those privacy-invasive methods in preference to providing more direct security. This is a case where general economic incentives skew what security methods are used. But those methods are very effective in compensating for cyber insecurity.

12 THE DECEPTIVELY TRANSPARENT BUT OPAQUE WORLD

The development of information technology does mean that nothing can be assured of staying secret. (The Manning, Martin, and Snowden security breaches at NSA cited above are only some of the most prominent examples.) There are just too many vulnerabilities in our systems and too many tools to capture and extract information, such as cameras in our cell phones and miniature cameras that are getting ever smaller and harder to detect. But neither can it be assumed that all relevant information will be available in forms that lead to action. The technique of "hiding in plain sight" was popularized by Edgar Allan Poe two centuries ago. Modern technology creates so much more information that this often works with minimal efforts at concealment, or even without any such effort. Even when information is known, it is often not known widely and is not known by people who might or should act on it. Just consider Dieselgate, where various groups had obtained measurements of emissions exceeding legal limits years before the scandal erupted. Or think of the Danish bank that laundered over $200 billion through a small Estonian branch over a few years—not to mention all the various sexual harassment cases that took ages to be noticed publicly.

In general, information that can be captured by information systems is becoming more detailed and far more extensive. But it is still limited in many ways. One of the most important ones is that human society is a messy affair and much that goes on is hard to codify precisely. In particular, tacit knowledge is crucial for individuals and organizations. Hence

even complete penetrations of computer systems of an organization are seldom sufficient to be able to replicate that organization's functioning. Studies have been carried out on the effects of East German espionage in West Germany. It was extremely effective at penetrating almost all targeted commercial organizations. But it allowed only a small narrowing in the performance gap between East and West German companies in the same industry. Especially when technology is advancing rapidly, the time to fully exploit information about current state of the art means that the intruders, who acquire the formal knowledge that is recorded, end up behind when they master those technologies.

As technology advances, the level of information that can be acquired increases, and so one might argue that the importance of tacit knowledge decreases. But that is very questionable. Systems are increasingly complicated, so it is harder to formally describe their functioning and their various failure modes and special features.

Further, modern technology allows for significant enhancements to the basic technique of "hiding in plain sight." Obfuscation techniques can be improved, and deployed much more widely and systematically, since we have increasing ability to create fake information. Looking forward, we are likely to see an arms race, with AI systems used to create "alternate realities" on one hand and to try to penetrate and deconstruct them on the other. The "post-truth" world is regarded as a danger, but it seems inevitable, and does have positive angles.

Note that there are many examples of primitive instances of such developments. The impenetrable legalese in the Terms of Service that users have to accept to use online services is a frequently encountered instance of what one recent paper referred to as "transparency [as] the new opacity." Much of what is done can be viewed as "speed bumps"—steps that are not guaranteed to be secure, but usually do offer some protection. An excellent example of that is provided by NDAs (nondisclosure agreements). Silicon Valley, which produces the bulk of the tools for eroding privacy and which often preaches the virtues of transparency, is full of them. Far from foolproof, they do serve to limit the spread and use of information.

13 THE VIRTUES OF MESSINESS

Lack of cybersecurity is universally regarded as just one aspect of the generally poor quality of our software, much of which is blamed on the "spaghetti code" nature of that software. But one should note that this poor quality also has positive aspects. Software piracy is not all that serious a problem, for example. Unpatched systems that are exposed on the Internet get easily penetrated. So frequent patching is required, and that means the software producer has to be in contact with systems running that code, and has a handle on illicit copies. Further, systems that are barely stable, and require constant upgrades to deal with bugs and improve functionality, cannot be easily adopted by competitors, which is another aspect of the tacit knowledge argument.

At a more mundane level, messiness of code, along with logging, is the primary reason digital forensics is as effective as it is. Attackers have difficulty covering up their traces. Much more can be done in this direction through intentional design.

Note that there are already successful examples of such approaches in the physical world. For example, color copiers generally have Machine Identification Codes (MICs), which leave a digital watermark on every page, identifying the printer and the date. (This case provides also another instance of successful "security through obscurity," since this technology was in wide commercial use for almost two decades and was not particularly secret, before it was widely publicized.)

A related approach is that of protecting consumer transactions by using diverse communication channels. Banks increasingly require confirmation of large and suspicious transactions through voice calls or texts—not as simple, quick, and cheap as letting Web entries go through, but capable of deployment in a flexible fashion, depending on the level of risk.

14 SPEED, REACH, AND COST FOR OFFENSE AND DEFENSE

At a very high level, information technologies have been revolutionary primarily because they offered quantum leaps in the three main measures of infrastructure effectiveness. They enabled actions or communications to be carried out much faster than was feasible before. They also allowed actions or communications to take place on a much wider scale. Finally, they did all of this at much lower cost.

These same advantages of information technologies, which led to so much progress in society, have also been attractive to criminals. Expert burglars could get into practically any dwelling, but it would usually take them some time to do so for every place. Automated probes can find and penetrate unpatched computers in seconds. Even an accomplished burglar needs some time, minutes or more typically hours, to rob a house. Hackers can commandeer thousands or even millions of computers in that time. Finally, all those attacks can be carried out at very low cost by hackers, who often don't even need much in the way of computers, as they can rely on ones they manage to seize control of.

But those same advantages of information technologies have also aided defense (just as happened with numerous earlier technologies). Defense can act much faster, as communication channels can be blocked, or software patched, far faster than physical locks could be changed. Centralized

defense teams can provide security for global organizations, without the need to station an armed guard at each location. And the costs are far lower than for physical protective measures.

Finally, there is that basic approach that was mentioned before: if it's too dangerous, don't use it. If high speed is a problem (as it is, as cryptocurrency enthusiasts keep discovering over and over and fail to learn from), slow things down. Don't allow large money transfers to occur until a day or two have passed, and there is a chance for monitoring systems (possibly ones involving loss of privacy) to collect and analyze data about the behavior of the entities involved. And so on.

These basic techniques underlie the usual approach taken by operators when faced with serious problems: bring down the network, repair (by reinstalling basic operating systems if necessary) all the machines that might be affected, and start bringing up functionality in sections of the network. That is how the now-ancient Morris worm infestation was dealt with. It is also how the collapse of campus network at a prestigious college was cured recently [2]. The ability of modern technology to operate in a decentralized fashion, with multiple ways of providing at least some basic functionality, is very helpful. As the report on that college's information systems debacle notes, when the basic network stopped functioning, the people involved "got creative." It's not something that one would undertake voluntarily, but it demonstrates the resilience of the system, and, among other things, makes it that much less attractive for attackers.

15 THE INCREASINGLY AMBIGUOUS NOTION OF SECURITY

Obfuscation, cited earlier, whether deliberate or accidental, will surely be an unavoidable and prominent feature of the "post-truth" world we are moving into. This world, full of information and misinformation, will create new challenges for security. To repeat the point made before, security is not the paramount goal by itself. But even beyond that dictum, we have to deal with the most fundamental questions of what security is and how it is to be provided. Increasingly it is not just about keeping out some well-defined "bad guys" out of the physical or cyber systems of an organization. The erosion of individual privacy tends to overshadow in the public mind the general explosion of information about organizations. Customers, suppliers, and partners legitimately possess an immense amount of information about any given enterprise. This information is being assembled in easily accessible format (for example, in the various customer relationship packages), which makes it easier to acquire and exploit. Therefore any enterprise is becoming less of a cohesive and isolated entity (physical or cyber) and more like a diaphanous web that overlaps other similar diaphanous webs. The problem of security in such a setting is then of managing the information flows to and from numerous other organizations, a much harder task than keeping out burglars or terrorists from a building.

In addition, security has always involved a very large dose of what Bruce Schneier has called "security theater." Security is often more about perceptions of security than about about any quantifiable and solidly established measures of security. Therefore security will increasingly overlap with public relations, and the generation of "spin."

16 CONCLUSIONS

This essay is a brief and very high level view of the cybersecurity area, in particular of how society has managed to thrive in spite of reliance on insecure information systems. The main conclusion is that, contrary to the public perception and many calls from prominent business and government leaders, we are not facing a crisis. This does not mean, though, that cybersecurity can be neglected, nor that all the effort that has been devoted to new security technologies has been wasted. Threats have been proliferating, and attackers have been getting more sophisticated. Hence new measures need to be developed and deployed. Firewalls are widely claimed to be becoming irrelevant. But they have been very useful in limiting threats over the last few decades. Now, though, we have to migrate to new approaches.

We do not know how to build secure systems of substantial complexity. But we can build very secure systems of limited functionality, and they can be deployed for specialized purposes, such as monitoring systems or ensuring integrity of backup systems, which are key to the ability to recover from hostile or accidental disasters.

We can also improve laws, regulations, and security standards. Cybersecurity is particularly rife with problems arising from the "tragedy of the commons" and negative externalities, and those problems can be mitigated. Microsoft dramatically improved the security of its products early in this century as a result of pressure from customers. Much more can be done this way. For example, it has been known that it is important to perform array bound checking, and how to do it, for half a century. It would not be too difficult to close that notorious hole that is key to numerous exploits.

The buffer overrun issue cited above brings up one of the main points of this essay, namely, that there are many ways to improve cybersecurity even without new inventions. What that means is that one has to be modest in expectations for anything truly novel. It may be a worthwhile goal to try for a "moonshot" or "silver bullet" technological solution in order to inspire the designers. But even if some dramatic breakthrough is achieved, it will still have to compete with a slew of other, more modest "Band-Aid" style approaches. So other factor than pure effectiveness, such as ease of use, may easily dominate and result in slow or no adoption.

This essay does suggest some contrarian ideas for increasing security. They are based on increasing complexity to enable many of the "speed bumps" that limit what attackers can do and help trace them. "Spaghetti code" has already been helpful and can be deployed in more systematic ways. In general, we should develop what Hilarie Orman has called a "theory of bandaids."

This essay does not claim that a "digital Pearl Harbor" will not take place. One, or more, almost surely will. But that has to be viewed in perspective. Given our inability to build secure system, such events may happen in any case. Further, their prospect has to be considered in comparison to the other threats we face. The issue is risk management, deciding how much resources to devote to various areas.

ACKNOWLEDGMENTS

The author thanks Ross Anderson, Steve Bellovin, Dorothy Denning, Ben Gaucherin, Balachander Krishnamurthy, Peter Neumann, Hilarie Orman, Walter Shaub, Robert Sloan, Bart Stuck, Phil Venables, Richard Warner, Bill Woodcock, and the editors of Ubiquity for their comments. Their providing comments should not be interpreted as any degree of endorsement of the thesis of this essay.

REFERENCES

[1] Geer, D. Risk Management Is Where the Money Is, *Risks Digest*, 20, 6, November 12, 1998. Available at https://catless.ncl.ac.uk/Risks/20/06. Accessed July 31, 2019.

[2] McKenzie, L. Amherst Students Incredulous About Going for Days Without Services They Consider Absolute Necessities, *Inside Higher Ed*, February 21, 2019. Avaiable at https://www.insidehighered.com/news/2019/02/21/almost-week-no-internet-amherst-college. Accessed July 31, 2019.

[3] New York Cyber Task Force. Building a Defensible Cyberspace, September 2017 report. Available at https://sipa.columbia.edu/ideas-lab/techpolicy/building-defensible-cyberspace. Accessed July 31, 2019.

PREFACE

This edition of *Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation* updates the previous two editions along two unfolding requirements: (1) a greater emphasis is placed on computer and network security, and (2) a lesser emphasis is placed on numerical and mathematical presentation. Enhancing computer security topics in most chapters and replacing Chapter 9 with "Hacking Social Networks" achieve the first change. The second emphasis is achieved by moving mathematical equations from the main body and placing them in several appendices. Additionally, qualitative frameworks are introduced that were absent in the second edition. The emphasis remains on rigorous methods, however. The author encourages the reader to undertake a rigorous approach to risk and resilience even without the benefits of mathematics.

From its inception, protection of infrastructure was a centerpiece of homeland security. After all, if you don't have food, water, energy, power, and communication, you don't have a country. The extreme vulnerability to accidental, weather-related, and human-instigated attacks on food, water, energy, power, transportation, and public health systems was understood years before 9/11, but nothing was done about it. Americans were not consciously aware of the criticality of their infrastructure until the devastation of September 11, 2001. Even then, public concern played second fiddle to the Global War on Terrorism. But the criticality of infrastructure has since moved to center stage in the public eye as America's roads and bridges decay, malware infects the Internet, transportation systems like air travel spread disease and terrorism, and the very financial underpinnings of modern society come increasingly under attack by hackers and unscrupulous speculators. Since 2001, the United States has experienced an historic series of system collapses ranging from the Middle Eastern wars to the financial debacle of 2008–2009. Some of the largest natural disasters in modern times have occurred. The Horizon oil spill in the Gulf of Mexico and Fukushima Daiichi tsunami/nuclear power disaster in Japan appear to be happening with greater frequency and consequence. Western civilization has taken on more debt than all of the previous empires combined, and most of the world's population is no more safe than it was in 2001. Unprecedented catastrophes continue to preoccupy us even as war appears to be subsiding. And almost all of these big events involve infrastructure. The lingering question is, why?

Our milieu is punctuated by historic crashes, collapses, and disasters followed by periods of calm that suddenly and without warning break out into a series of crashes, collapses, and disasters once again. Why is this? The answer lies deep in modern life's complexity. That is, modern society runs on complex systems that never existed prior to the last century. For example, the concept of a big business did not exist before the railroads in the 1870s, and the concept of a big system did not exist before the construction of the power grid, food and water supply chains, and global travel networks. In only a few short decades, people and communications infrastructure have become interdependent with just about all other systems as the Internet connects more people, places, and things into one big complex system. Modernity means connectivity and connectivity means complexity. As it turns out, complexity is the main source of risk and fragility in critical infrastructure and key resource (CIKR) systems. So, the short answer to the question of why collapses appear to be getting bigger and more frequent is simply complexity.

The level of complexity of systems we depend on is the root cause of extreme calamity—not some change in the climate, absence of forethought, or ineptitude. After all,

buildings can be made to withstand hurricanes and earthquakes, and roads and bridges can be built to withstand ever more severe damage. Rather than blame our failing systems on weather and terrorists, deeper analysis suggests that collapse is built into infrastructure itself because of its structural complexity. Risk is found in the way systems are "wired together" and operated. Fragility—the opposite of resilience—is another symptom of complexity.

According to Bak's theory of self-organization, modern society is responsible for the fragility of the very infrastructure it depends on for survival. Modern designers are clever—they have optimized systems so completely that infrastructure systems have no wasteful surge capacity. Roads and bridges are built to be cost-efficient, not necessarily resilient. Water and power systems owned by both public and private corporations are built to return a profit, not necessarily to last for centuries. Power lines are much cheaper to build above ground than below, which means they become useless when a superstorm such as Sandy strikes the Eastern seaboard. Public health institutions such as hospitals have no need for extra beds or 30-day power supplies, because such a lengthy outage could never happen and the cost of resilience is simply too high. Supply chains are made efficient by centralizing distribution centers without concern for possible terrorist attacks that can take out an entire sector. Truly unthinkable events are shuffled off to insurance companies to worry about. What is the risk of an unthinkable incident of unbounded size, but near zero probability of occurring?

In the summer of 2002, when the original material for this book and its first edition was being collected, I did not understand the connection between critical infrastructure and complex systems. It seemed to me that critical infrastructure systems were largely mechanical and electrical machines. If the components worked, the entire system also worked. If something went wrong, only that something was affected. I had no concept of "system," or "complexity." I thought the study of infrastructure was roughly equivalent to the study of processes and mechanical inputs and outputs. If students understood how these machines worked, and how they might be damaged, they could do something to protect them. System resilience is the sum of resilient components, and hardening targets against a variety of hazards removes risk. Thus, in 2003 I began writing the first edition of this book as a kind of "engineering lite" textbook.

The first edition covered the structure of water, power, energy, and Internet sectors—mostly from an engineering point of view. Because my students were nontechnically trained, I included some historical background and lightly tossed in organizational things like how a sector is regulated or how regulation impacts infrastructure. From 2003 to 2006 my students struggled with the math and engineering concepts and graciously provided useful feedback. After more than two years of class testing, John Wiley & Sons agreed to publish the first edition. But by then I had the nagging feeling that I missed the mark. Why did complex systems like the power grid fail more often than they should? Why do epidemics like SARS explode onto the international scene and then vanish just as quickly? Why do terrorists attack patterns look a lot like accidents? These system questions cannot be answered by engineering or organizational psychology methods of analysis, because complex system behavior is "more than the behavior of its parts." I came to realize that infrastructure is not designed and built, but instead, it is an emergent process that evolves. Most infrastructure sectors defined by the Department of Homeland Security are hostage to a number of hidden forces. As a result, most infrastructure has emerged as a complex system subject to unpredictable behavior when stressed. They fail in unexpected ways when under stress and sometimes fail even when perturbations are small. The nonlinearity of cause and effect captured my interest, which led me to consider complexity and the new science of networks as a means of understanding fragility.

I came to realize that the biggest threat to infrastructure systems was their topology—their architecture. Vulnerability to natural or human-made collapse is built in to these systems. The secret to understanding them is buried within their very structure—largely defined by connections and interdependencies. For the most part, critical infrastructure is critical because of the way it is put together. And construction of most sectors is largely accidental or emergent. The commercial air transportation system is fragile because of important hubs (airports), the power grid is weakened by substations and transmission lines that handle more than their share of "connectivity," and the monoculture Internet is prone to malware because it has so many connections that malicious code may travel from one side of the globe to the other with ease. It is structure—in the form of network connectivity—that makes critical infrastructure vulnerable to collapse.

I learned that Perrow's normal accident theory explained why small incidents sometimes spread and magnify in intensity until consequences are catastrophic. The cause of this spread, according to Perrow, is hidden coupling—invisible links inherent in complex social and mechanical systems. Perrow's breakthrough theory laid the blame for catastrophic failure on the system itself. A spark may start a fire, but it is fuel, in the form of kindling, that spreads the flames and builds consequence. Per Bak—one of the founders of complexity theory—reinforced the idea that widespread collapse is inherent in the system itself. Bak went a step further than Perrow, however, and postulated the theory of self-organization. In Bak's theory, complex systems become more fragile as they age, due to a number of factors. The most common factor simply being gradual restructuring as a system attempts to optimize performance. Bak called this self-organizing criticality (SOC).

SOC can be measured in almost all critical infrastructure systems using a mathematical quantity called the spectral radius. This may seem like an exotic quantity, but it is simply a measure of connectivity in a system. A power grid, supply chain, transportation system, the Internet, and most every infrastructure can be represented as a network. Nodes can be substations, depots, warehouses, Internet service providers, bridges, and so on and links represent their connectivity. The pattern or "wiring diagram" of the network model is called the infrastructure's topology. Topology has a fingerprint quantified as the spectral radius. Interestingly, spectral radius increases as the density of connections increases. It also increases as hubs form. Hubs are components that are overly connected through links to other components, such as a busy airport or central office of the telephone company.

Another metric found to be extremely useful is the fractal dimension of a system. This quantity simply measures fragility of a complex system by relating the likelihood of cascade failures (as described by Perrow in his normal accident theory) to spectral radius. One can think of component vulnerability as the likelihood of propagating a fault from one asset to another and fractal dimension as the propensity of an entire system to cascade. Higher propensity to collapse becomes a measure of resilience. Resilience goes down as vulnerability and spectral radius goes up. This relationship applies to all infrastructure sectors that can be represented as a network of connected components. If we want to increase infrastructure resilience, we must decrease spectral radius, component vulnerability, or both.

By introducing formal network theory, infrastructures can be modeled and studied in the abstract. This makes it possible to understand and measure risk and resilience of infrastructure at a conceptual level. Infrastructure resilience hinges on the structure of the system—not just its component's weaknesses. Inherent weakness (vulnerability due to weak components and self-organized structure) can then be addressed on a system scale rather than a component or single-asset scale. Safe and secure policies can be designed to address the inherent risk of system collapse, instead of patchwork guessing. By measuring spectral radius and fractal dimension of various infrastructure systems, we can provide policy-makers with scientific tools upon which to make policy. Does changing a regulation reduce spectral radius? Does hardening of one asset make other assets more likely to fail?

Like its predecessors, this edition first develops a general theory of risk, resilience, and redundancy and then applies the general theory to individual sectors. After an introductory chapter and three chapters on the theoretical foundations of risk and resiliency, the focus shifts to structural (architectural) properties of communications, Internet, information technology, SCADA, water, energy, power, public health, transportation, supply chains (shipping), and banking systems. Each chapter describes a sector and then applies network science and complexity metrics to an actual or hypothetical CIKR system. This should provide the reader with general tools that he or she can apply to other systems. It is a unified theory approach to the topic.

Unlike its predecessors, this edition includes qualitative risk and resilience frameworks that use checklists and qualitative rankings to measure risk and resilience. The NIST Cybersecurity Framework (CSF) is a qualitative framework for evaluating computer security risk and resilience. The Community Fragility Framework proposed by Lori Hodges is featured because it is based on complexity theory—one of the fundamental foundations of critical infrastructure protection. While these are less rigorous and certainly less quantitative, they often cover a broader array of factors than quantitative frameworks.

This edition stands on the shoulders of the second edition and feedback from educators that have used previous editions in their classrooms. Many colleagues contributed, but I would especially like to thank Paul Stockton of Sonecon, Inc., Susan Ginsburg of Criticality Sciences, Waleed al Mannai of the Kingdom of Bahrain, Harry Mayer of Health and Human Services, Brady Downs and Eric Taquechel of the US Coast Guard, Michael Larranaga, Massoud Amin, Chris Bellavita, Rodrigo Nieto-Gomez, Richard Bergin, Glen Woodbury, Lori Hodges, Mathem Liotine, Richard H. Martin, Bernard A. Jones, Kristine Twomey, Jackie L. Deloplaine, Robert Crane, and Mike Walker. As always, the errors and flaws in this book are my responsibility alone.

tedglewis@icloud.com TED G. LEWIS
Monterey, CA, USA
June 2019

HOW TO USE THIS BOOK

What is critical about critical infrastructure? What is the biggest threat to infrastructure? Who owns and operates the Internet? Is terrorism a high-risk threat? Can cybersecurity ever be achieved? What is the most resilient infrastructure in America? These and other questions are addressed in this book. They are the most often asked questions of students of safety and security studies. They are also practical questions asked by fire protection, law enforcement, public administration, urban planning, criminology, political science, and homeland security practitioners. The answers are organized into 18 chapters roughly divided into three parts: Part I, Origins of Homeland Security and Critical Infrastructure Protection Policy (Chapter 1); Part II, Theory and Foundations (Chapters 2–4); and Part III, Individual Sectors (Chapters 5–17). In addition, there is a strategy chapter 18, and there are four appendices containing supplemental material on probability, risk, spectral radius, tragedy of the commons, and encryption algorithms, and a glossary of terms—for the extra-curious and mathematically prepared reader.

This material has been used in a 12-week hybrid course entitled "Critical Infrastructure Protection: Vulnerability and Analysis," taught for over a decade at the Center for Homeland Defense and Security (CHDS) in Monterey, California. CHDS is a great resource for additional teaching materials and references. Most materials including supplements to this book are available for free at www.CHDS.us or from the author at tedglewis@icloud.com.

There are two general categories of readers for this course: the policy student/practitioner with minimal background in science, technology, engineering, and mathematics (STEM) and the STEM student/practitioner. The former is advised to skip over sections containing mathematical equations without loss of understanding of the concepts. The latter are advised to read the appendices as well as the body of material in each chapter.

The main concepts of critical infrastructure protection are covered in Chapters 2–4. First, Chapter 2 surveys risk analysis—theory and its practical analysis. Appendix A contains more detailed and mathematical treatment of the topic, including an introduction to Bayesian belief networks. Chapter 3 surveys the predominant theories of catastrophe—Perrow's normal accident theory and Bak's punctuated equilibrium theory. In addition, this chapter incorporates the tragedy of the commons, paradox of enrichment, Gause's law of competitive exclusion, and the paradox of redundancy—properties of complex systems discovered, for the most part, by biologists. Critical infrastructure is treated much like a living organism in these sections, because like a living and breathing organism, infrastructure evolves, adapts, and changes when stressed.

Chapter 4 is the capstone of the three chapters dealing with theory. It surveys network science and shows how to apply the fundamental concepts of complex networks to infrastructure analysis. This is where self-organized criticality, fractal dimension, and spectral radius are introduced and illustrated using examples taken from a variety of sectors. Chapter 4 formally defines the concept of a hub, betweener, and blocking node—three important tools used to improve resiliency in any complex system. An appendix explains the fundamental equation of resilience and defines the relationship between fractal dimension, spectral radius, and component vulnerability—an equation that applies to nearly all infrastructure systems.

Chapters 5–17 apply the techniques and models of Chapters 2–4 to the communications, information technology, SCADA,

water, energy, power, public health, transportation, shipping, and banking sectors. Chapter 18 suggests general strategies for protecting infrastructure. Most of the analysis is original and provides insights not previously reported in the literature. For example, a handful of blocking nodes are critical to the continuity of operation of the Internet.

Finally, a number of supporting materials are available from the publisher and author. An instructor's manual containing answers to the exercises and PowerPoint slide decks containing lectures are available from Wiley.com.

ABOUT THE COMPANION WEBSITE

This book is accompanied by a companion website:

www.wiley.com/go/Lewis/CriticalInfrastructure_3e

The website includes Instructor's Guide and Instructors Slides.

1

ORIGINS OF CRITICAL INFRASTRUCTURE PROTECTION

What is the motivation for studying *critical infrastructure protection* (CIP)? What are the central issues that need to be addressed in order to create a meaningful strategy for dealing with threats against infrastructure? We begin by tracing the development of *CIP* over several decades and noting that it has evolved through at least eight phases: from initial awareness to combating terrorism, emphasis on natural disaster response, an early definitional phase, a public–private cooperation phase, a federalism versus states phase, a resilience awareness phase, a risk-based decision-making phase, and after massive computer security breaches and the failure of government to "wake up to" the realities of computer and network exploits at both misdemeanor and warlike levels, the cybersecurity phase.

CIP is a multifaceted topic because it cuts across many disciplines and jurisdictions. It cuts vertically across federal, state, local, and tribal political boundaries, and it cuts horizontally across public and private organizations. It has a variety of policy issues at one extreme and a diverse set of scientific and engineering issues at the other extreme. The most glaring example of this is the electric power grid, which is pulled in many different directions by political, social, engineering, and public–private forces. The rapid emergence of online e-commerce, social networks, and misinformation campaigns also raise political, social, and engineering issues broadly classified as cybersecurity threats and exploits. The topics in this book touch on all of these, at architectural and policy levels, by applying complexity theory and network science to the practical problem of securing critical infrastructure and key resources (CIKR).

One of the most difficult tasks of protecting critical infrastructure (CI) is the problem of deciding who is responsible for what across these political and organizational lines. While policy at the Department of Homeland Security (DHS) offices in Washington, DC, may advocate an all-hazard risk-informed decision-making process and encourage community action, actual operational and organizational processes at the state and local level may be entirely different due to a number of factors. Federalism and top-down policy-making may look good on paper, but actual implementation at the local level often lacks jurisdictional clarity, required expertise, willpower, or all three. For example, what is the role of public safety responders such as firefighters and law enforcement officers when something goes wrong with a gas pipeline, electrical power fails during a storm, or hackers exploit the Internet in a city without cybersecurity expertise?

There remain gaps in knowledge, jurisdictional clarity, and organizational fitness—challenges this book attempts to address—in the emerging field of CIP. As this chapter illustrates, the field is still evolving. Some issues are being resolved, while others are still in the early stages of their evolution. The field has matured, circa 2019, after decades of slow but steady maturation, such as follows:

- *Recognition*: No such field of study existed prior to the mid-1900s. Although awareness of the importance of infrastructure began in 1962 with the Cuban Missile Crisis, nearly 30 years passed before the term *critical infrastructure protection* was defined. Throughout

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

these 30 years, the roles and responsibilities of governmental agencies as well as the definition of CIP changed as the field evolved. Nonetheless, much remained to be resolved in this initial phase.

- *Natural disaster recovery*: In the beginning, CIP was nearly identical to *consequence management*—recovery from disasters such as floods, hurricanes, and earthquakes. The Stafford Act[1] established the Federal Emergency Management Agency (FEMA)—a federal agency dedicated to recovery after a flood, hurricane, earthquake, tornado, and so on. Terrorism was not a factor in CIP in the beginning. It would take a decade of attacks before CIP was linked with terrorism in the United States. But a focus on terrorists—human-caused incidents—soon faded as natural disasters occurred more often than terrorist attacks, and headlines focused the public's attention on billion-dollar natural disasters.
- *Definitional phase*: The term "critical infrastructure" did not exist before the 1990s. There was no definition of CIP, and infrastructure was taken for granted. The public was confident that freshwater always flowed from faucets and electric light switches always produced light. The terrorist attacks of 9/11 changed all that, of course, even though the earliest definition of CIP was made in 1997. Then, from 1997 through 2003, the identification of CI sectors expanded from eight to 13 sectors plus 5 *key assets*, expanded again to 18 sectors and key resources (KR), and then consolidated into 16 CIKR sectors in 2013. Today it is difficult to identify sectors of the national economy that are *not* critical; however, this book attempts to define criticality in a rigorous and operational way.
- *Public–private cooperation*: The role of the private sector in CIP was slow to take root until the late 1990s. But so many CIKR assets are in the hands of corporations—not local, state, or federal government—that it is difficult to separate public versus private assets. Public safety and health, law enforcement, and emergency response are largely a function of local government, but energy, power, communications, and commercial air travel are largely a function of the private sector. Water and key assets such as dams fall somewhere in between. Who should respond when something happens to these systems? Even today, the federal government and private sector owners of infrastructure are not clear on their respective roles and responsibilities with respect to CIP, although the role of government in protecting systems of all types has narrowed over the decades. Nonetheless, when a small business in mid-America is hacked by a teenager running scripts downloaded from the dark web, it is not clear who is responsible for the protecting the small business from the availability of the script, dark web, teenager, or Internet service provider.
- *Federalism*: Because terrorists attack at the local level, the solution to the problem must also come from the local level—states, cities, and tribes. The future of homeland security rests in the hands of local governments, and yet the local level lacks sufficient technology, skills, and funding to cope with global terrorism, computer criminals, or major catastrophes. Superstorm Sandy, Fukushima Daiichi power plant disaster, the Horizon Gulf Oil spill, Russian hackers from the Internet Research Agency, and major droughts throughout the Midwest routinely outstrip local government's capacity to deal with catastrophic events—both physical and virtual. Federal–state–local–tribal federalism does not seem to be able to cope with CIKR events spanning political boundaries or that are so consequential that local authorities are overwhelmed.
- *Resilience*: By the mid-2000s it became obvious that asset hardening and 100% security of the vast CIKR sectors was an impossible and impractical goal of CIP. CIKR systems are too big, too complex, and too expensive to protect in their entirety. Thus, the field entered a period of reevaluation and government agencies began to focus on *resiliency* rather than absolute security.[2] Although the definition of risk and resilience went through many iterations, the concept of a resilient society began to take root as an objective of CIP. However, like the term *risk*, the term *resiliency* still lacks a standard definition, making the application of resilience-informed decision-making difficult and often ineffective. A plethora of frameworks such as the DHS risk management, the National Institute of Standards and Technology Cybersecurity Framework (NIST CSF), and Hodges Community Fragility model appeared at about this time as early attempts to formalize resilience.
- *Risk-informed decision-making*: Ten years after the horrific terrorist attacks of September 11, 2001 (9/11), the notion of resilience remained a laudable goal but difficult to measure. Therefore, the field of CIP entered a more quantifiable phase loosely called *risk-informed decision-making*.[3] During this phase, a variety of methods and practices emerged to quantify risk and

[1]The Stafford Act is a 1988 amended version of the Disaster Relief Act of 1974.

[2] From PPD-21, resiliency is defined as "the ability to prepare for and adapt to changing conditions and withstand and recover rapidly from disruptions. Resilience includes the ability to withstand and recover from deliberate attacks, accidents, or naturally occurring threats or incidents."

[3]From the DHS Risk Lexicon, risk-informed decision-making is "determination of a course of action predicated on the assessment of risk, the expected impact of that course of action on that risk, as well as other relevant factors."

resilience and to measure the return on investment (ROI) for a variety of techniques ranging from target hardening, public–private partnerships (PPP), regulation of chemicals, and others to rigorous methods of assessing risk and resilience. Risk-informed decision-making seeks to prioritize investments in infrastructure on the basis of quantitative risk assessment. The DHS and FEMA released a semiquantitative measure of risk to assist local agencies quantify risk of CIKR within their agencies.

- *Cybersecurity and infrastructure*: Mounting losses due to computer security breaches both commercially and within government began to be counted as viable threats to national security. Perhaps the initial awareness occurred with the weaponized Stuxnet exploit, but it is more likely that Russian meddling and misinformation campaigns by Russia during the 2016 US presidential election was the lightning rod that prompted action by President Trump in 2018 to re-organize the DHS's CIP bureaucracy via the *Cybersecurity and Infrastructure Security Agency Act of 2018* (CISA). CISA elevated computer security as a major threat to CIKR in particular and government-owned and government-operated computer and network systems in general. In 2019 we entered the cybersecurity phase of CIKR evolution.

There is little reason to believe the cybersecurity and infrastructure phase is a final stage of evolution because of unforeseen threats ahead. Modern society is in a headlong dash toward even greater global connectivity and adoption of Promethean technologies such as 5G, artificial intelligence, cryptocurrencies, quantum computing, quantum communications, and elevated consequences of global climate change. Greek god Prometheus gave fire to humans—perhaps the first technology with both good and evil applications. The Promethean challenge of our age is to enjoy the benefits of technology while also controlling it. This challenge has yet to be met in the field of CIP.

1.1 RECOGNITION

Prior to the dramatic and horrific attacks of September 11, 2001 (9/11), the US public had little awareness of terrorism or how it could impact them personally. Attacks on the homeland were something that happened in other countries—not the United States. But a growing number of "national security emergencies" culminating in 9/11 exposed terrorism for what it is—a challenge to the security of the people of the United States. Even before 9/11 however, a few policy-makers were busy formulating various strategies and policies that culminated in a national strategy for homeland security. A major part of this national strategy involved CIP—the protection of basic infrastructure sectors such as water, power, telecommunications, health and medical services, the Internet, and transportation systems. The early work of this small group peaked in the late 1990s, which marks the origins of what we now call *homeland security*. During this same time, CI and CIP emerged as a key element of homeland security.

Although CIP was defined and recognized as a major component of national security rather late in the game (1996), it really began with the creation of the National Communications System (NCS) in 1963 after communications problems between the United States and the Soviet Union threatened to interfere with negotiations during the Cuban Missile Crisis[4]:

> In October [1962], President John F. Kennedy, on national television, revealed that the Soviets had placed nuclear missiles in Cuba. As a result of this aggressive action, he ordered quarantine on all offensive military equipment under shipment to Cuba until the Soviets removed their weapons. … For nearly a week, the Nation was transfixed by the actions of Soviet Premier Nikita Khrushchev and President Kennedy. During this time, ineffective communications were hampering the efforts of the leaders to reach a compromise. Without the ability to share critical information with each other using fax, e-mail, or secure telephones such as we have today, Premier Khrushchev and President Kennedy negotiated through written letters. Generally, Washington and Moscow cabled these letters via their embassies. As the crisis continued, hours passed between the time one world leader wrote a letter and the other received it. Tensions heightened. On October 27 and 28, when urgency in communications became paramount, Premier Khrushchev bypassed the standard communication channels and broadcast his letters over Radio Moscow.
>
> Following the crisis, President Kennedy, acting on a National Security Council recommendation, signed a Presidential memorandum establishing the NCS. The new system's objective was "to provide necessary communications for the Federal Government under all conditions ranging from a normal situation to national emergencies and international crises, including nuclear attack."
>
> At its inception on August 21, 1963, the NCS was a planning forum composed of six Federal agencies. Thirty-five years later, it is a vital institution comprising 23 member organizations that ensure NS/EP (National Security/Emergency Preparedness) telecommunications across a wide spectrum of crises and emergencies. … During the 1980s and 1990s, the NCS expanded its focus to develop Government wide NS/EP procedures and enhancements to the Nation's public networks and information infrastructures.

The role of the communications infrastructure grew more important as the United States entered the information age. In 1978, two communications regulatory agencies (Department

[4] http://www.ncs.gov/about.html

of Commerce's Office of Telecommunications and the White House Office of Telecommunications) were combined into the National Telecommunications and Information Administration (NTIA) by Executive Order 12046. NTIA handled the process of selling spectrum to telephone, radio, and TV networks. It also has the distinction of being the federal agency that oversaw the commercialization of the Internet in 1998–1999. The NCS was formally assigned responsibility for the telecommunications infrastructure in 1984 by Executive Order 12472.

In 1982 President Reagan established the National Security Telecommunications Advisory Committee (NSTAC) by Executive Order 12382. This important presidential advisory body is made up of the CEOs of the major telecommunications companies.

NSTAC is perhaps the first organization to advise a president on CIP.

The NCS and the NSTAC were the first CI agencies within the US government. Twenty years would pass before the term *critical infrastructure* would be defined and the entire US population would become aware of its importance in their daily lives. The DHS absorbed NCS in February 2003, but the NSTAC still reports to the President of the United States.

1.2 NATURAL DISASTER RECOVERY

While the NCS and NSTAC were active throughout the 1970s and 1980s, responses to disasters—both human caused and natural—were still on the back burner as far as CIP was concerned. The FEMA was created in 1978–1979 to respond to hurricanes and earthquakes.[5] Soon after its creation, FEMA was assigned the (temporary) responsibility of responding to terrorist attacks by Executive Order 12148 in 1979[6]:

> All functions vested in the President that have been delegated or assigned to the Defense Civil Preparedness Agency, Department of Defense, are transferred or reassigned to the Director of the Federal Emergency Management Agency.
>
> All functions vested in the President that have been delegated or assigned to the Federal Disaster Assistance Administration, Department of Housing and Urban Development, are transferred or reassigned to the Director of the Federal Emergency Management Agency, including any of those functions re-delegated or reassigned to the Department of Commerce with respect to assistance to communities in the development of readiness plans for severe weather-related emergencies.
>
> All functions vested in the President that have been delegated or assigned to the Federal Preparedness Agency, General Services Administration, are transferred or reassigned to the Director of the Federal Emergency Management Agency.
>
> All functions vested in the President by the Earthquake Hazards Reduction Act of 1977 (42 U.S.C. 7701 *et seq*.), including those functions performed by the Office of Science and Technology Policy, are delegated, transferred, or reassigned to the Director of the Federal Emergency Management Agency…. *For purposes of this Order, "civil emergency" means any accidental, natural, man-caused, or wartime emergency or threat thereof, which causes or may cause substantial injury or harm to the population or substantial damage to or loss of property*.

FEMA was confronted by perhaps the first major terrorist attack on US soil in Oregon in 1984. Members of the politico-religious commune founded by Bhagwan Shree Rajneesh[7] attempted to influence a political election by poisoning voters with salmonella:[8]

> In a bizarre plot to take over local government, followers of Bhagwan Shree Rajneesh poisoned salad bars in 10 restaurants in The Dalles in 1984, sickening 751 people with salmonella bacteria. Forty-five of whom were hospitalized. It is still the largest germ warfare attack in U.S. history. The cult reproduced the salmonella strain and slipped it into salad dressings, fruits, vegetables and coffee creamers at the restaurants. They also were suspected of trying to kill a Wasco County executive by spiking his water with a mysterious substance. Later, Jefferson County District Attorney Michael Sullivan also became ill after leaving a cup of coffee unattended while Rajneeshees lurked around the courthouse.
>
> Eventually, Ma Anand Sheela, personal secretary of the Bhagwan, was accused of attempted murder, conspiracy, arson, and other crimes and disowned by the Bhagwan. Convicted of the charges against her, she spent 29 months in federal prison, then moved to Switzerland.[9]

The salmonella incident in Oregon was an attack on one of many infrastructure sectors identified as critical over the past decade: *agriculture*. But in 1984 there was no generally accepted definition of *infrastructure*, nor any recognition of what sectors belonged to the list of national *CI*.

The importance of infrastructure began to dawn on the federal government when in 1988 President Reagan issued Executive Order 12656. This order alludes to "essential

[5]Presidential Reorganization Plan No. 3 issued by President Carter in 1978 established the Federal Emergency Management Agency (FEMA), which went into effect on April 1, 1979.

[6]http://www.archives.gov/federal_register/codification/executive_order/12148.html

[7]http://www.religioustolerance.org/rajneesh.htm

[8]"The group settled on the 65,000 acre 'Big Muddy Ranch' near Antelope, Oregon, which his *sannyasins* had bought for six million dollars. The ranch was renamed *Rajneeshpuram* ('Essence of Rajneesh'). This 'small, desolate valley twelve miles from Antelope, Oregon was transformed into a thriving town of 3,000 residents, with a 4,500 foot paved airstrip, a 44 acre reservoir, an 88,000 square foot meeting hall…'" http://www.clui.org/clui_4_1/lotl/lotlv10/rajneesh.html

[9]http://home.att.net/~meditation/bioterrorist.html

resources" and places responsibility for their protection in the hands of federal departments:

> The head of each Federal department and agency, within assigned areas of responsibility shall:
>
> **Sec. 204.** *Protection of Essential Resources and Facilities*.
>
> (1) Identify facilities and resources, both government and private, essential to the national defense and national welfare, and assess their vulnerabilities and develop strategies, plans, and programs to provide for the security of such facilities and resources, and to avoid or minimize disruptions of essential services during any national security emergency;
> (2) Participate in interagency activities to assess the relative importance of various facilities and resources to essential military and civilian needs and to integrate preparedness and response strategies and procedures;
> (3) Maintain a capability to assess promptly the effect of attack and other disruptions during national security emergencies.

This executive order contains a number of objectives that remain problematic even today. It calls for identification of public and private facilities that are essential to national welfare—a task that remains unfulfilled today, as political and socioeconomic forces complicate the definition of "essential" and "national welfare." A bridge in one county may be considered essential by voters in that county, but not essential in an objective sense, because of alternative routes. Moreover, when limited resources are considered and there is funding for only one bridge, objective selection of which bridge is saved or repaired quickly enters the political realm instead of the rational realm.

Part two of President Reagan's executive order calls for interagency cooperation to address military and civilian needs. When a severe emergency such as a devastating superstorm or terrorist attack happens, however, interagency cooperation often vanishes and the military takes over. Civil–military relations theoretically means that the military takes orders from civilians, but in practice, only the military has the capacity to deal with major catastrophes. This inequality between the authority of local law enforcement agencies and the readiness of federal troops is revealed over and over again whenever major incidents such as Hurricane Katrina and New Orleans spin out of control.

Finally, the third part of the executive order remains problematic because state and local agencies often do not or cannot afford to maintain capabilities to meet the need. For example, a smallpox outbreak in Manhattan—a population of 8 million—would quickly overwhelm public health and safety agencies in New York. The state and local authorities would have to maintain 40,000 trained emergency responders to head off the spread of smallpox. Forest fires in California quickly overwhelmed firefighters in 2018 and illustrated the importance of interagency and interregional (reciprocal) response agreements in the Department of Interior.

1.3 DEFINITIONAL PHASE

Even in the early 1990s the trend toward greater awareness of human-made and natural disasters was subtle—it had not yet reached a point where it was of national concern. But by 1993–1995 the rate and severity of acts of terror, for example, was increasing and becoming more alarming to the federal government. The 1993 attack on the World Trade Center by Ramzi Yousef, the acts and eventual capture of the Unabomber (1995), the devastating attack on the Federal Building in Oklahoma City, Oklahoma (1995), and the sarin gas attack in a Tokyo subway in 1995 suggested a trend. Acts of violence by nongovernmental organizations (NGOs) were increasing, and as a by-product, raising the level of public awareness. Soon these acts would be attributed to terrorists and move from the back to the front page of the media. Within a short 5–6 years, response to unlawful terrorism would become known as the *Global War on Terrorism* (GWOT) and reached a threshold that deserved national attention.

During this definitional phase, the importance of infrastructure to the safety and security of the US population began to take shape. But the threat was still confined to human-initiated acts of terror. One of the earliest concerns was the fragility and vulnerability of the systems we depend on daily, such as roads, bridges, stadiums, schools, and office buildings. These facilities accommodate many people and yet they are completely open and unprotected. The communications systems and the energy and power systems that run cities and enable modern society to function were also open and unprotected. The emergency response systems and public health services taken for granted for decades were suddenly exposed as poorly prepared. Modern life depended on them, and yet, these essential systems were vulnerable to attacks by both humans and Mother Nature.

The modern origin of homeland security and one of its pillars, CIP, can be placed somewhere between 1993 and late 1995. In fact, 1995 is a reasonable start date because of the flurry of activity aimed at protecting national infrastructure and key assets after 1995. Presidential Decision Directive 39 (PDD-39) issued by President Clinton in 1995 set the stage for what was to come—a new federal Department of Homeland Security. PDD-39 essentially declared war on terrorists[10]:

> It is the policy of the United States to deter, defeat and respond vigorously to all terrorist attacks on our territory and against our citizens, or facilities, whether they occur domestically, in international waters or airspace or on

[10]http://www.fas.org/irp/offdocs/pdd39.htm

> foreign territory. The United States regards all such terrorism as a potential threat to national security as well as a criminal act and will apply all appropriate means to combat it. In doing so, the U.S. shall pursue vigorously efforts to deter and preempt, apprehend and prosecute, or assist other governments to prosecute, individuals who perpetrate or plan to perpetrate such attacks.
>
> We shall work closely with friendly governments in carrying out our counterterrorism policy and will support Allied and friendly governments in combating terrorist threats against them. Furthermore, the United States shall seek to identify groups or states that sponsor or support such terrorists, isolate them and extract a heavy price for their actions. It is the policy of the United States not to make concessions to terrorists.

The criticality of national infrastructure and associated key assets became an important issue when President Clinton issued Executive Order 13010 (EO-13010) in 1996. This executive order established a Presidential Commission on Critical Infrastructure Protection (PCCIP). The commission was chaired by Robert Marsh and subsequently became known as the *Marsh Report* [1]. It defined *critical infrastructure* in terms of "energy, banking and finance, transportation, vital human services, and telecommunications." The Marsh Report was the first publication to use the term critical infrastructure and has become one of the foundational documents of CIP.

The Marsh Report and EO-13010 provided the first formal definition of *infrastructure* as "a network of independent, mostly privately-owned, man-made systems that function collaboratively and synergistically to produce and distribute a continuous flow of essential goods and services." And *critical infrastructure* is "an infrastructure so vital that its incapacity or destruction would have a debilitating impact on our defense and national security."

According to EO-13010,[11]

> Certain national infrastructures are so vital that their incapacity or destruction would have a debilitating impact on the defense or economic security of the United States. These critical infrastructures include telecommunications, electrical power systems, gas and oil storage and transportation, banking and finance, transportation, water supply systems, emergency services (including medical, police, fire, and rescue), and continuity of government. Threats to these critical infrastructures fall into two categories: physical threats to tangible property ("physical threats"), and threats of electronic, radio frequency, or computer-based attacks on the information or communications components that control critical infrastructures ("cyber threats"). Because many of these critical infrastructures are owned and operated by the private sector, it is essential that the government and private sector work together to develop a strategy for protecting them and assuring their continued operation.

[11]http://www.fas.org/irp/offdocs/eo13010.htm

The work of the PCCIP resulted in PDD-63 (Presidential Decision Directive of 1998), which defined CI more specifically and identified eight basic sectors, listed in Table 1.1. According to PDD-63,

> Critical infrastructures are those physical and cyber-based systems essential to the minimum operations of the economy and government. They include, but are not limited to, telecommunications, energy, banking and finance, transportation, water systems and emergency services, both governmental and private.[12]

Table 1.1 identifies the sectors initially defined by PDD-63 in 1998 and also identifies the sector-specific agency (SSA) responsible at the federal level. SSAs can be any government agency responsible for carrying out the various CIP missions (Page 50 in Ref. [2]):

- Leads, integrates, and coordinates the execution of the National Infrastructure Protection Plan (NIPP), in part by acting as a central clearinghouse for the information sharing, reporting, and coordination activities of the individual sector governance structures.
- Facilitates the development and ongoing support of governance and coordination structures or models.
- Facilitates NIPP revisions and updates using a comprehensive national review process.
- Ensures that effective policies, approaches, guidelines, and methodologies regarding partner coordination are developed and disseminated to enable the SSAs and other partners to carry out NIPP responsibilities.
- Facilitates the development of risk, risk-informed, and criticality-based assessments and prioritized lists of CIKR.
- Facilitates the sharing of CIKR prioritization and protection-related best practices and lessons learned.
- Facilitates participation in preparedness activities, planning, readiness exercises, and public awareness efforts.
- Ensures cross-sectoral coordination with the SSAs to avoid conflicting guidance, duplicative requirements, and reporting.

The definition of CI in PDD-63 went through rapid evolution and expansion after the attacks of 9/11. The Office of the President of the United States released the National Strategy for Homeland Security in July 2002 and then rapidly followed up with an expansion of the definition of CI sectors in February 2003 with the release of the National

[12]http://www.fas.org/irp/offdocs/pdd/pdd-63.htm

TABLE 1.1 The basic critical infrastructure sectors (8) defined by PDD-63 (1998)

Sector	Description	Sector-specific agency
1. Banking and finance	Banking and stock markets	Treasury
2. Emergency law enforcement services	Justice/FBI	Justice
3. Emergency services	Emergency fire and continuity of government	FEMA
4. Energy	Electric power, gas and oil production and storage	Energy
5. Information and communications	Telecommunications and the Internet	Commerce
6. Public health services	Public health, surveillance, laboratory services, and personal health services	HHS
7. Transportation	Aviation, highways, mass transit, rail, pipelines, shipping	Transportation
8. Water supply	Water and its distribution	Environmental Protection Agency

Strategy for the Physical Protection of Critical Infrastructures and Key Assets.[13]

According to the 2003 strategy document, the objectives of CIP include:

- Identifying and assuring the protection of those infrastructures and assets that we deem most critical in terms of national-level public health and safety, governance, economic and national security, and public confidence consequences.
- Providing timely warning and assuring the protection of those infrastructures and assets that face a specific, imminent threat.
- Assuring the protection of other infrastructures and assets that may become terrorist targets over time by pursuing specific initiatives and enabling a collaborative environment in which federal, state, and local governments and the private sector can better protect the infrastructures and assets they control.

In addition to the list of sectors shown in Table 1.2, the 2003 National Strategy lists five KR:

- National monuments and icons
- Nuclear power plants
- Dams
- Government facilities
- Commercial key assets

TABLE 1.2 CIKR (14) as of 2003

Sector	Sector-specific agency
Agriculture	Dept. of Agriculture
Food	
• Meat and poultry	Dept. of Agriculture
• All other food products	Dept. of Health and Human Services
• Water	Environmental Protection Agency (EPA)
• Public health	Dept. of HHS
• Emergency services	Dept. of Homeland Security
Government	
• Continuity of government	Dept. of Homeland Security
• Continuity of operations	All departments and agencies
• Defense industrial base	DOD
• Information and telecommunications	Dept. of Homeland Security
• Energy	Dept. of Energy
• Transportation	Dept. of Homeland Security (TSA)
• Banking and finance	Dept. of the Treasury
• Chemical industry and hazardous materials	EPA
Postal and shipping	Dept. of Homeland Security
Nat'l monuments and icons	Dept. of the Interior

1998 was a year of ramping up counterterrorism programs. Major initiatives besides PDD-62 (Countering Terrorism), PDD-63 (Critical Infrastructure Protection), and PDD-67 (Continuity of Government) were the creation of a variety of programs:

- National Infrastructure Protection Center established in the Department of Justice.
- Chemical Safety Board formed.
- National Domestic Preparedness Office created in the Department of Justice.
- Critical Infrastructure Analysis Office (CIAO) established.
- Counter-Terror Coordination Unit in National Security Council formed.

[13]The National Strategy for the Protection of Critical Infrastructures and Key Assets, February 2003. Department of Homeland Security. http://www.dhs.gov

- Congress earmarks $17 M for Special Equipment and Training Grants.
- Attorney General announces creation of National Domestic Prep. Office (NDPO).

1.4 PUBLIC–PRIVATE COOPERATION

By 1999 some experts believed that most infrastructure in the United States was owned by the private sector—not government. The Internet had just been commercialized in 1998 and the communications and electrical power sectors were in the process of being deregulated. Control of most public utilities was in the hands of corporations, and according to Table 1.1, it appeared that the private sector owned or operated most infrastructure considered "critical."[14] Thus, in 1999 President Clinton established the National Infrastructure Assurance Council (NIAC) to bring industry and government closer together. According to Executive Order 13130, NIAC was established to facilitate the partnership through the Public Sector Information Sharing and Analysis Centers (PS-ISAC)[15]:

> By the authority vested in me as President by the Constitution and the laws of the United States of America, including the Federal Advisory Committee Act, as amended (5 U.S.C. App.), and in order to support a coordinated effort by both government and private sector entities to address threats to our Nation's critical infrastructure, it is hereby ordered as follows:
>
> **Section 1.** *Establishment*.
>
> (a) There is established the National Infrastructure Assurance Council (NIAC). The NIAC shall be composed of not more than 30 members appointed by the President. The members of the NIAC shall be selected from the private sector, including private sector entities representing the critical infrastructures identified in Executive Order 13010, and from State and local government. The members of the NIAC shall have expertise relevant to the functions of the NIAC and shall not be full-time officials or employees of the executive branch of the Federal Government.
>
> (b) The President shall designate a Chairperson and Vice-Chairperson from among the members of the NIAC.
>
> (c) The National Coordinator for Security, Infrastructure Protection and Counter-Terrorism at the National Security Council (National Coordinator) will serve as the Executive Director of the NIAC.
>
> (d) The Senior Director for Critical Infrastructure Protection at the National Security Council will serve as the NIAC's liaison to other agencies.
>
> (e) Individuals appointed by the President will serve for a period of 2 years. Service shall be limited to no more than 3 consecutive terms.
>
> **Section 2.** *Functions*.
>
> (a) The NIAC will meet periodically to:
>
> (1) enhance the partnership of the public and private sectors in protecting our critical infrastructure and provide reports on this issue to the President as appropriate;
>
> (2) propose and develop ways to encourage private industry to perform periodic risk assessments of critical processes, including information and telecommunications systems; and
>
> (3) monitor the development of Private Sector Information Sharing and Analysis Centers (PS-ISACs) and provide recommendations to the National Coordinator and the National Economic Council on how these organizations can best foster improved cooperation among the PS-ISACs, the National Infrastructure Protection Center (NIPC), and other Federal Government entities.
>
> (b) The NIAC will report to the President through the Assistant to the President for National Security Affairs, who shall assure appropriate coordination with the Assistant to the President for Economic Policy.
>
> (c) The NIAC will advise the lead agencies with critical infrastructure responsibilities, sector coordinators, the NIPC, the PS-ISACs and the National Coordinator on the subjects of the NIAC's function in whatever manner the Chair of the NIAC, the National Coordinator, and the head of the affected entity deem appropriate.

1.5 FEDERALISM: WHOLE OF GOVERNMENT

The National Strategy document of 2003 declares that homeland security and CIP in particular are "whole of government" responsibilities. "Homeland security, particularly in the context of critical infrastructure and key asset protection, is a shared responsibility that cannot be accomplished by the federal government alone. It requires coordinated action on the part of federal, state, local, and tribal governments; the private sector; and concerned citizens across the country."[16]

But in practice, the strategy places most of the power—and all of the funding—in the hands of the federal government. For example, all SSAs are federal government agencies. The federal government assumed this responsibility even before the creation of the DHS in 2003. The President's Critical Infrastructure Protection Board (PCIPB) was one of the earliest federal government agencies created as a consequence of 9/11. It was followed by a flurry of

[14]The source of this claim has never been found, but a popular meme of the time was that the private sector owned or operated 85% of the critical infrastructure listed in Table 1.1.

[15]http://www.archives.gov/federal_register/executive_orders/1999.html#13130

[16]http://www.fas.org/irp/offdocs/pdd/pdd-63.htm

additional government bureaucracies created to counterterrorism and natural disasters—incidents that appeared to be rising exponentially.

By Executive Order (EO) 13231 (October 2001), President Bush created the President's PCIPB, with primary responsibility to develop policies to protect the information infrastructure of the federal government. EO 13231 recognized the growing importance of the telecommunications and Internet infrastructure as well as its interdependency with other sectors. Without information systems, the US federal government could not continue to operate in the event of an attack:

> Consistent with the responsibilities noted in section 4 of this order, the Board shall recommend policies and coordinate programs for protecting information systems for critical infrastructure, including emergency preparedness communications, and the physical assets that support such systems.

In 2002 President Bush signed the Homeland Security Bill, establishing the new DHS. It began operation in February 2003 and incorporated 22 agencies that were scattered throughout the federal bureaucracy. This included the NCS, CIAO, and the Department of Justice Office of Domestic Preparedness, along with a number of other large agencies such as the TSA, INS, Border Patrol, and Coast Guard. Protection of CI continued to expand and become one of the major responsibilities of the DHS.

Presidential Directive HSPD-5 (February 2003) and its companion, *HSPD-8* (December 2003), authorized the Secretary of DHS "to prevent, prepare for, respond to, and recover from terrorist attacks, major disasters, and other emergencies"[17] In December 2003 President Bush replaced PDD-63 with HSPD-7 (Homeland Security Presidential Directive No. 7). It rewrote the list of sectors and SSAs responsible (see Table 1.3).

Unfortunately, HSPD-7 sectors and KR departed from the list given by the National Strategy and clouded the issue of which department or agency was responsible for energy, power, and the information and telecommunications sector. The list of CIKR in Table 1.3 was short-lived.

Indeed, HSPD-7 does *not* specify who is responsible for several of the sectors previously identified as "critical." It appears that HSPD-7 was written to address infighting among departments and agencies that may have felt left out of the National Strategy. Alternatively, the purpose of HSPD-7 may have been to include departments and agencies that have expertise in fields such as cyber, chemical, and nuclear security. For whatever reason, HSPD-7 leaves some responsibilities unspecified and spreads others across multiple departments.

[17]HSPD-5 (2003).

TABLE 1.3 CIKR (16) and responsibilities as defined by HSPD-7

Sector	Sector-specific agency
Agriculture/food (meat, poultry, eggs)	Department of Agriculture
Public health/food (other than meat, poultry, eggs)	Department of Health and Human Services
Drinking water and treatment systems	Environmental Protection Agency
Energy (production, storage, distribution of gas, oil, and electric power, except for commercial nuclear power facilities)	Department of Energy
Nuclear power plants	Department of Homeland Security and Nuclear Regulatory Commission and Department of Energy
Banking and finance	Department of the Treasury
Defense industrial base	Department of Defense
Cybersecurity	Department of Commerce and Department of Homeland Security
Chemical	Not specified
Transportation systems, including mass transit, aviation, maritime, ground/surface, and rail and pipeline systems	Department of Transportation and Department of Homeland Security
Emergency services	Not specified
Postal and shipping	Not specified
National monuments	Department of the Interior
Key assets: dams, government facilities, and commercial facilities	Not specified

For the first time, HSPD-7 declared that it is impractical to protect everything and focused effort on major incidents—ones that cause mass casualties comparable to the effects of using weapons of mass destruction:

> While it is not possible to protect or eliminate the vulnerability of all critical infrastructure and key resources throughout the country, strategic improvements in security can make it more difficult for attacks to succeed and can lessen the impact of attacks that may occur. In addition to strategic security enhancements, tactical security improvements can be rapidly implemented to deter, mitigate, or neutralize potential attacks… Consistent with this directive, the [DHS] Secretary will identify, prioritize, and coordinate the

> protection of critical infrastructure and key resources with an emphasis on critical infrastructure and key resources that could be exploited to *cause catastrophic health effects or mass casualties* comparable to those from the use of a *weapon of mass destruction*. [3]

By 2009, the number of sectors and KR had expanded even more, culminating in 18 CIKR: *critical manufacturing* was added and information technology and communications were separated into two sectors [2]. In less than a decade, the number of CIKR expanded from 8 to 18. At this pace, CIP would embrace just about every aspect of society, from communications, power, and healthcare to the food we eat, water we drink, and work we do. If CIP embraces nearly everything, perhaps it means nothing. What then is the main goal of CIP?

HSPD-5 and HSPD-8 were expanded by President Obama on March 30, 2011, to strengthen "… the security and resilience of the United States through systematic preparation for the threats that pose the greatest risk to the security of the Nation, including acts of terrorism, cyber attacks, pandemics, and catastrophic natural disasters."[18] President Obama pared down the number of CIKR in HSPD-7 to 16 sectors and KR in PPD-21 (2013) (see Table 1.4). Postal and shipping was folded into transportation and national monuments and icons were removed. In addition, the SSAs responsible for each CIKR were sharpened with more authority given to the DHS. Thus, the long-term definition of CI was established, but it emphasized physical assets more than cyber assets. This changed in 2018.

TABLE 1.4 CIKR as defined by PPD-21 (2013)

Sector	Sector-specific agency
Chemical	Department of Homeland Security
Commercial facilities	Department of Homeland Security
Communications	Department of Homeland Security
Critical manufacturing	Department of Homeland Security
Dams	Department of Homeland Security
Defense industrial base	Department of Defense
Emergency services	Department of Homeland Security
Energy	Department of Energy
Financial services	Department of the Treasury
Food and agriculture	US Department of Agriculture and Department of Health and Human Services
Government facilities	Department of Homeland Security and General Services Administration
Healthcare and public health	Department of Health and Human Services
Information technology	Department of Homeland Security
Nuclear reactors, materials, and waste	Department of Homeland Security
Transportation systems	Department of Homeland Security and Department of Transportation
Water and wastewater systems	Environmental Protection Agency

[18]PPD-8 (2011).

A series of events precipitated a major realignment within the DHS in late 2018. Major information security breaches of National Security Agency (NSA) documents by Edward Snowden (1983) in 2013, followed by WikiLeaks releasing emails and documents exfiltrated from the Democratic National Committee during the 2016 US presidential election campaign, and misinformation campaigns waged by the Russian Internet Research Agency attempting to influence the 2016 US presidential election precipitated a renewed focus on cyber as well as physical security within the DHS. The 2018 CISA legislation created the CISA organization as shown in Figure 1.1.

On November 16, 2018, President Trump signed into law the *Cybersecurity and Infrastructure Security Agency Act* of 2018. This legislation emphasized cybersecurity for the first time and replaced the National Protection and Programs Directorate (NPPD) with the Cybersecurity and Infrastructure Security Agency also referred to as CISA:

> **CISA's Cybersecurity Division** works with government and private sector customers to ensure the security and resilience of the Nation's cyber infrastructure. The division includes the National Cybersecurity Communications Integration Center (NCCIC).
>
> The **Emergency Communications Division** enhances public safety interoperable communications at all levels of government, providing training, coordination, tools and guidance to help partners across the country develop their emergency communications capabilities.
>
> The **Infrastructure Security Division** coordinates security and resilience efforts using trusted partnerships across the private and public sectors, and delivers training, technical assistance, and assessments to federal stakeholders as well as to infrastructure owners and operators nationwide.
>
> The **National Risk Management Center** (NRMC) works to identify and address the most significant risks to our nation's critical infrastructure.

The CISA leads the national effort to defend CI against the threats of today while working with partners across all levels of government and in the private sector to secure against the evolving risks of tomorrow.

1.6 RISE OF THE FRAMEWORK

A precursor to the risk-informed decision-making phase of DHS was the rise of the framework. A framework is a particular set of rules, ideas, or beliefs used to structure

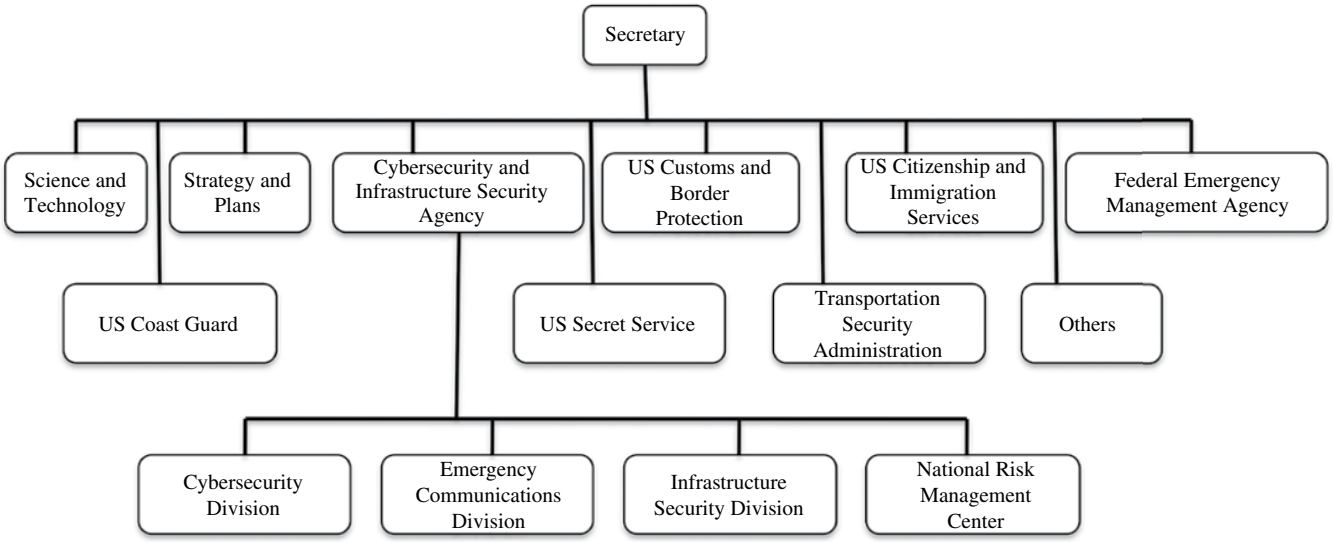


FIGURE 1.1 The structure of the cybersecurity and infrastructure protection offices within the Department of Homeland Security as of 2019 is focused on cybersecurity, emergency communications, infrastructure security, and risk management.

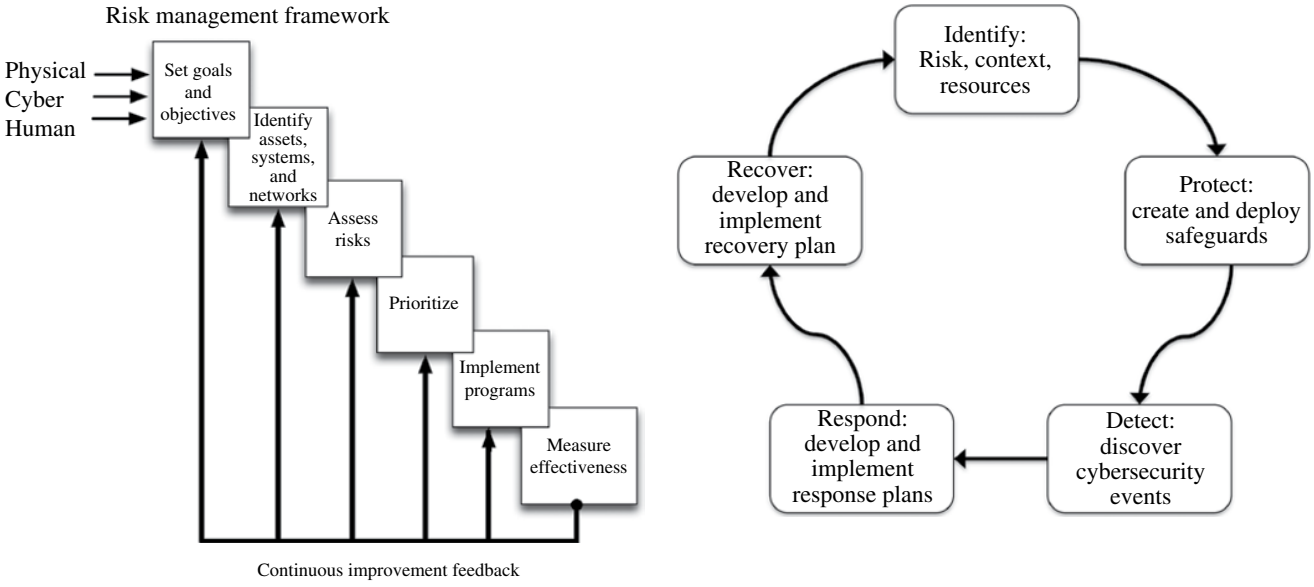


FIGURE 1.2 Two frameworks for qualitative risk management—one for physical assets and the other for computer and network exploits. (a) An early DHS risk management framework for critical infrastructure. (b) The NIST Cybersecurity Framework (CSF) for defending against computer and network exploits.

decision-making. Figure 1.2a is an early framework for risk-informed decision-making within DHS. Figure 1.2b is a specialized framework for evaluating risk and making qualitative risk-based decision-making within the cybersecurity realm.

A number of competing and sometimes overlapping frameworks exist for organizing efforts to protect CI systems. These frameworks can be roughly categorized as **political**, **qualitative**, **quantitative**, and **regulatory/legal**. This book leans toward the quantitative frameworks, but it is important to note that others exist in both theory and practice. A short description of each type is given here with longer descriptions of quantitative frameworks given throughout this book.

Political frameworks have existed since the beginning of government's recognition of CIKR as a federal, state, local, and tribal responsibility. For example, the first allocation of resources formula to combat terrorist attacks on CIKR was based on a mix of population and politics. Each region was

allocated funding regardless of the need. Emergency response facilities such as firefighting equipment were funded regardless of risk or the likelihood of threats. Politically, this made sense because large population centers are where the voters are. However, the embarrassing reality is that some of the most critical assets such as the largest nuclear power plant in the nation are located far from population centers. Threats are more likely to be high where CI assets are high impact, regardless of population or risk.

Qualitative frameworks such as the NIST CSF began to appear as checklists and recommendations to owners and operators of industrial control systems, power grids, and water system SCADA. Executive Order 13636 (EO-13636), *Improving Critical Infrastructure Cybersecurity* (February 2013), and the *Cybersecurity Enhancement Act* of 2014 (CEA) established the role of the NIST in identifying and developing cybersecurity risk frameworks (CSF) for use by CI owners and operators. NIST claims the CSF is "a prioritized, flexible, repeatable, performance- based, and cost-effective approach, including information security measures and controls that may be voluntarily adopted by owners and operators of critical infrastructure to help them identify, assess, and manage cyber risks."

Version 1.1 (April 2018) of the CSF prescribes a five-step process along with checklists of recommended practices (see Fig. 1.2b):

- **Identify**—Develop an organizational understanding to manage cybersecurity risk to systems, people, assets, data, and capabilities. Understanding the business context, the resources that support critical functions, and the related cybersecurity risks enables an organization to focus and prioritize its efforts, consistent with its risk management strategy and business needs.
- **Protect**—Develop and implement appropriate safeguards to ensure delivery of critical services. This step supports the ability to limit or contain the impact of a potential cybersecurity event.
- **Detect**—Develop and implement appropriate activities to identify the occurrence of a cybersecurity event.
- **Respond**—Support the ability to contain the impact of a potential cybersecurity incident.
- **Recover**—Develop and implement appropriate activities to maintain plans for resilience and to restore any capabilities or services that were impaired due to a cybersecurity incident.

The framework is a hierarchical checklist for computer system owners and operators. For example, the **Protect** step might be further decomposed into sub-steps:

User credential verification, revocation, and device authorization.

Physical access permissions.

Remote access permissions.

Network configuration and integrity.

Personnel awareness and training.

Data security—at rest and in transit.

Data capacity assurance.

Separation of development systems from operational systems.

Configuration change controls.

Backup maintenance.

Response and recovery plans are tested.

Vulnerability management plan in place.

Audit records implemented and maintained

Removable media is protected.

Communications and control networks are protected.

Fail-safe, load-balancing mechanisms implemented for resilience.

While NIST claims CSF is a risk-based approach to managing cybersecurity risk, the framework does not define risk or resilience and offers no specific risk assessment methodology or model. Users are left to their own definition of risk and resilience, which is often qualitative rather than quantitative.

Regulatory/legal frameworks follow a similar process diagram of continual improvement. However, for most of its history, DHS has deferred to other agencies when it comes to tying CIKR security to regulations and legal requirements. Generally, regulation has been applied to safety and environmental protections more than security. However, this remains a largely untapped potential source of CIKR protection. For example, the vulnerability of the communications sector is heavily dependent on regulation and the 1996 Telecommunications Act, which created the highly critical carrier hotels and concentrated assets vulnerable to both physical and cyber attacks. This topic is covered in more detail in Chapters 5–8.

The final category of framework is the one emphasized in this book—quantitative—the use of formulas and equations to quantify risk and resilience in what has become known as risk-informed decision-making. A preview of this approach is given here, but the remainder of this book focuses on quantitative measures as much as possible.

1.7 IMPLEMENTING A RISK STRATEGY

The overall strategy of CIP was set by 2012 with PDD-21, but implementation remained a challenge. Policy dictated a vertically integrated effort from federal–state–local and tribal governments and a horizontally integrated effort across public and private organizations. Government was supposed to cooperate, and the private sector was supposed to help the

public sector. But what does this mean? What was each party supposed to do?

Roles and responsibilities could not be aligned vertically or horizontally without operational definitions of objectives. Broadly, the objectives of CIP were impractical as stated by policy. Specifically, infrastructure is too vast, complex, and expensive to protect everything, and expertise among governmental agencies is nonexistent. This called for a narrower definition of objectives and operational definitions of goals, for example, government had to define what is critical in a CI, and both public and private parties had to agree upon metrics for prioritizing projects. Before CIP policy can be implemented, goals and objectives must be defined rigorously enough to implement them.

Policy stated the obvious—protect infrastructure from hazards such as terrorists, storms, earthquakes, and so on. Protection included both hardening and response when something bad happens. Funding was inadequate to protect everything, so implementation depended on prioritization of CI assets, which in turn depended on the definition of *criticality*. Two approaches were initially attempted. The first prioritization strategy was called *risk-informed* and the second was called *resilience-informed*. Risk-informed decision-making means applying risk assessments to prioritize funding of projects to harden CI assets. Resilience-informed decision-making means applying various methods to enhance the resilience of infrastructure assets. Rather than hardening assets, resilience-informed decision-making attempts to make assets adaptable and anti-fragile. Both approaches have their strengths and weaknesses.

1.7.1 Risk-Informed Decision-Making

The fundamental question posed by a risk-informed strategy is this: given limited resources of the federal government, how should resources (funding) be allocated to reduce risk? How should priorities be set? Once again, we turn to the NIPP 2009 for guidance:

> ***Risk***. *The potential for an unwanted outcome resulting from an incident, event, or occurrence, as determined by its likelihood and the associated consequences*.
>
> ***Risk-Informed Decision-making***. *The determination of a course of action predicated on the assessment of risk, the expected impact of that course of action on that risk, and other relevant factors*.
>
> ***Risk Management Framework***. *A planning methodology that outlines the process for setting goals and objectives; identifying assets, systems, and networks; assessing risks; prioritizing and implementing protection programs and resiliency strategies; measuring performance; and taking corrective action*.

The era of risk-informed decision-making evolved slowly from politically motivated allocation of resources to the quantifiable and measurable six-step process described in Figure 1.1. Instead of dividing funding according to pressures from politicians, risk-informed decision-making allocates funding according to the likelihood of a high-consequence event. Risk is defined in different ways by different SSAs, but given a rigorous definition of risk, agencies can allocate funds according to their impact on risk reduction. The risk-informed strategy follows a risk assessment process such as the following (see Fig. 1.1):

1. Set goals and objectives: Objectives may range from reduction of consequences to elimination of risk, increasing resiliency, and risk minimization. A risk-informed decision-making emphasizes risk reduction, but may also consider additional objectives such as sociopolitical benefits to a community.
2. Identify assets, systems, and networks: Single assets such as a building, bridge, computer, or ports are easy to identify, but most CIKR are part of a complex system. For example, there are numerous assets in a water system—pipes, pumps, treatment plants, and reservoirs. Thus, drinking water is a system containing many assets typically connected together in some fashion. Generally, these systems are modeled as a network of nodes and links: nodes representing pumps, treatment plants, and reservoirs and links representing pipes.

   Assess risks: Risks can be calculated in a variety of ways. A multi-criteria risk assessment is a spreadsheet containing risk factors and numerical ratings for each factor. A probabilistic risk assessment (PRA) approach is more exacting: the simplest form is $R = TVC$, where T is threat as defined by the probability of a human attacker, V is the vulnerability of the asset or system to a given threat, and C is consequence. V is a conditional probability that a given threat will succeed if attempted. C is consequence measured in a meaningful unit such as dollars, casualties, or economic damage. See Appendix B for mathematical details.

   For natural disasters and accidents, a different risk equation is used: $R = \mathrm{E}(c)C$, where $\mathrm{E}(c)$ is the probability of a hazardous event obtained from historical data and C is consequence as before. Hazard probabilities are known for floods, hurricanes, earthquakes, and tornadoes. For example, the famous Gutenberg–Richter scale for measuring the intensity of earthquakes is actually a probability distribution that relates the likelihood of an earthquake to its intensity c. An earthquake of 8 is 1 million times more intense than an earthquake of 4 on the Gutenberg–Richter scale. But the probability E(4) of a magnitude 4 earthquake is 10^{-4} and the probability E(8) of a magnitude 8 earthquake is 10^{-8}—10,000 times less likely.

Risk assessment becomes more complicated when analyzing a complex adaptive system such as a power grid, human population subject to an infectious disease, or large and complex transportation system. When such CIKR systems are assessed for risk, we must consider nonlinear effects, feedback loops, and a variety of factors. These are discussed in subsequent chapters.

Prioritize: CIKR are typically so large and expensive that it is necessary to identify the most critical assets of vital importance. This requires prioritization—a lengthy topic in itself. Simple prioritization in a risk-informed decision-making setting might be to rank assets according to risk. The highest-risk assets are allocated resources first. But this has limitations because the cost to reduce risk by 1% may differ greatly from one asset to another. If the goal is to reduce overall risk, then it may be better to reduce the most cost-effective risks first. In this case, reducing risk of the highest-risk assets may not be cost-effective.

A number of prioritization schemes should be considered. For example, consider highest-consequence, most-vulnerable, highest-risk, highest-return-on-investment, and highest-increase-in-resiliency schemes, depending on the goals and objectives of the risk management framework. A variety of optimization techniques may be applied to this step, because in the end, prioritization is a resource allocation problem that answers the question, "what is the best use of resources to minimize or maximize the objective?"

3. Implement programs: A typical assessment of CIKR produces a recommendation. For example, the assessment may advise the community to secure its drinking water system, repair bridges, or buy backup transformers for the local power grid. Each of these actions takes investment of resources—most often in the form of funding. The outputs from the previous step (Prioritize) are used to guide these investments.

   Measure effectiveness: Finally, the effectiveness of the implementation program needs to be measured and feed back into subsequent assessments. A simple measure is ROI. For example, if the objective is to reduce risk, ROI is obtained by calculating the difference in risk before and after program implementation and dividing by the amount of investment:

$$ROI = \frac{Risk(before) - Risk(after)}{\$Investment}$$

The risk-informed strategy is labor intensive, because all assets must be evaluated and numerical values of T, V, and C estimated. These measurements may number in the thousands, and because it involves probabilities, they may be inaccurate. Furthermore, the results of risk assessment may not satisfy sociopolitical objectives such as addressing assets critical to one segment of the population at the expense of assets in other segments of the population. How does one choose between protecting the drinking water system in one part of town versus the hospital in another part of town?

1.7.2 Resilience-Informed Decision-Making

Almost immediately upon the formation of the new DHS it became clear that CIKR assets numbered in the millions (see Table 1.5) (Page 50 in Ref. [2]). The vastness of single sectors makes it impossible to protect everything. When multiplied by the large number of sectors and key assets, the challenge became insurmountable without some kind of prioritization. Furthermore, the concept of "100% security" began to vanish and be replaced by an elusive concept—*resilience*. Instead of an unyielding goal of 100% security, resilience was an intangible property of CIKR somewhere between absolute security and absolute vulnerability. Instead of a secure infrastructure, a resilient infrastructure was able to bounce back after being attacked or damaged by a storm, earthquake, and so on.

The February 2003 National Strategy document contained the word *resilience* three times. The NIPP 2009 document mentions resilience 15 times. The 2013 PPD-21 directive from President Obama incorporates resilience in its title and

TABLE 1.5 Selection of CIKR assets

Assets in a select subset of CIKR
1,912,000 farms
87,000 food-processing plants
1,800 federal reservoirs
1,600 municipal wastewater facilities
5,800 registered hospitals
87,000 US localities
250,000 firms in 215 distinct industries
2 billion miles of cable
2,800 power plants
300,000 producing sites
5,000 public airports
120,000 miles of major railroads
590,000 highway bridges
2 million miles of pipelines
300 inland/costal ports
500 major urban public transit operators
26,600 FDIC insured financial institutions
66,000 chemical plants
137 million delivery sites
5,800 historic buildings
104 commercial nuclear power plants
80,000 dams
3,000 government-owned/government-operated facilities
460 skyscrapers

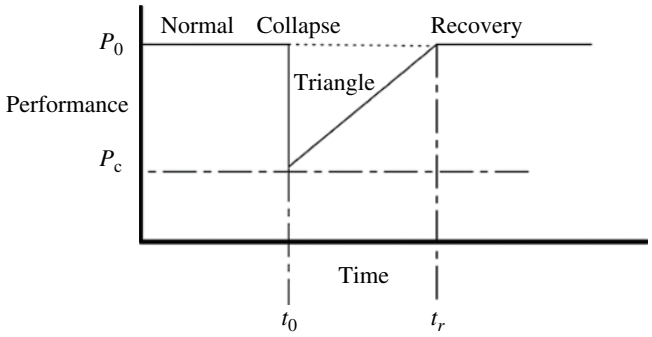


FIGURE 1.3 A resilience triangle is formed by a collapse followed by recovery.

uses the word 44 times.[19] By 2013 the focus of CIKR had shifted from counterterrorism and all-hazard preparedness to building resilience into both infrastructure and the population. The era of resilient infrastructure began, and terrorism, all-hazard response, and weapons of mass destruction faded into the background.

Unfortunately, a variety of qualitative definitions of resilience make it difficult to measure and apply. Vurgin et al. surveyed the concept of resilience in infrastructure systems and offered a number of definitions [4]. Generally, resilience is a property of a system—not a single asset:

> Given the occurrence of a particular disruptive event (or set of events), the resilience of a *system* to that event (or events) is the ability to efficiently *reduce both the magnitude and duration* of the deviation from targeted system performance levels.[20]

Of course, this definition is difficult to put into practice, because it lacks quantifiable specifics. Bruneau et al. proposed a measurable and operational model of resilience as shown pictorially in Figure 1.3 and mathematically modeled in Appendix B. Damage to a system in the form of magnitude and duration is represented by a triangular area notched out of a performance-versus-time diagram shown in Figure 1.3. The resilience triangle represents loss due to a drop in performance followed by a recovery period that eventually restores the system to its previous level of performance.

The difference between full performance and diminished performance represented by the resilience triangle defines the system's resilience. Smaller triangular areas represent greater resilience. The size of the triangular area is reduced, by reducing (1) recovery time, (2) precipitous drop in performance, or (3) both. In addition, the likelihood of a precipitous drop in performance increases the frequency of collapses over time. Thus, reducing the size of the resilience triangle increases resilience:

1. Speedup recovery: $(t_r - t_0)$
2. Reduce performance drop: $(P_0 - P_c)$
3. Decrease the probability of failure, V

This definition suffices for single assets such as buildings, bridges, Internet servers, power plants, and pipelines, but it is inadequate to quantify the resilience of complex interdependent systems such as the power grid, communications network, or an entire municipal water system. However, this metric quantifies the qualitative definition of resilience proposed in the NIPP 2009:

> ***Resilience****: The ability to resist, absorb, recover from, or successfully adapt to adversity or a change in conditions*. (Page 111 in Ref. [2])

But the resilience triangle model does not address resistance, absorption, adaptation, and recovery factors loosely defined by the NIPP. How does a CIKR resist, absorb, or recover from adversity? How is the ability to resist, absorb, or adapt to adversity measured? These complex properties are addressed by a *complex adaptive systems* model of CIKR described in more detail in Chapters 2–4.

1.7.3 Prevention or Response?

Both risk- and resilience-informed strategies beg the question "How much should be devoted to response versus prevention?" When it comes to CIKR protection, is prevention the best use of resources, or should money be spent mainly in response? In a prevention-only strategy, resources are applied to deter and prevent damage. A response-only strategy invests in response capability, such as emergency management services, law enforcement and firefighting capacity, and so on.

One way to answer to this question is to classify hazards according to their risk levels—low, high, or even complex. Figure 1.4 illustrates the difference between high- and low-risk hazards. The risk profile curve of Figure 1.4 shows how risk can increase without bound versus consequence or approach zero after a temporary increase. The profile of a

[19]Presidential Policy Directive 21—Critical Infrastructure Security and Resilience.

[20]The source of this claim has never been found, but a popular meme of the time was that the private sector owned or operated 85% of the critical infrastructure listed in Table 1.1.

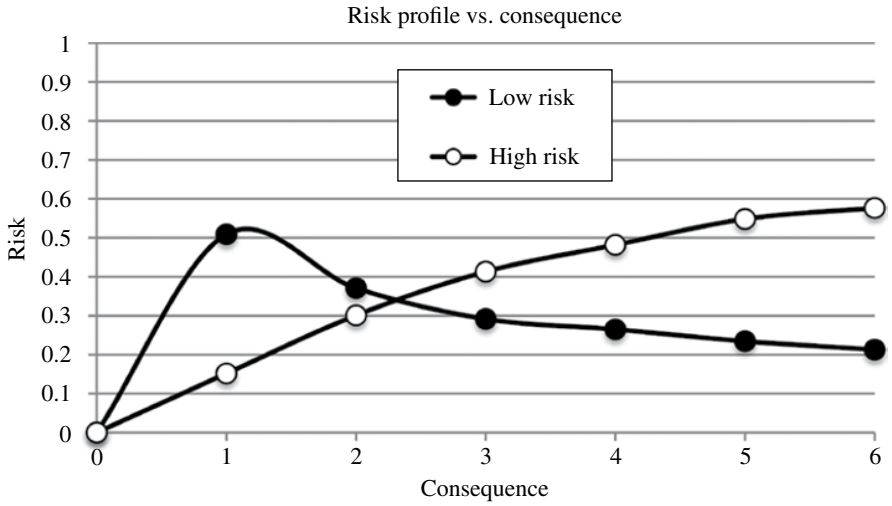


FIGURE 1.4 Some hazards are low risk and some are high risk. Risk increases for high consequences when a hazard is high risk, and the opposite is true for low-risk hazards.

low-risk hazard approaches zero as consequence approaches infinity. The profile of a high-risk hazard approaches infinity.

One of the persistently unresolved CIKR security issues is the question of how many resources should be applied to prevention versus response: is the strategy biased more toward response as the National Strategy seems to suggest, or does it provide just as much support for prevention? What should the balance between prevention and response be?

An argument for a greater emphasis on prevention is that prevention is cheaper than suffering mass casualties, economic damage, psychological damage, or damage to National pride. But 100% prevention is impossible. Some terrorist acts will always escape detection and natural disasters like hurricanes cannot be avoided. Still, roads, buildings, and power lines can be designed to withstand almost anything—for a price.

Table 1.6 lists some high- and low-risk hazards, based on their risk profiles. Note that some consequences are measured in deaths, some in financial loss, and others in impacted area. Consequence can be expressed in a number of different units. Furthermore, risk due to an earthquake is considered low, when measured in land area, but high when measured in deaths.

Figure 1.4 suggests a different risk-informed strategy for low- versus high-risk hazards. For example, the financial risk of small city fires is considered high risk. Therefore, strict building codes and inspections are called for to prevent them. The opposite strategy might apply to low-risk hazards such as terrorism and airline accidents. More resources should be applied to response. Thus, the best risk-informed strategy might depend on the profile of the hazard:

Prevention vs. Response: *Apply more resources to prevention of high-risk hazards and more resources to response to low-risk hazards.*

1.8 ANALYSIS

The evolution of CIP continues to expand and encompass a wider array of challenges. From a focus on terrorism, the homeland security enterprise has grown to encompass cybersecurity, response to natural disasters and climate change, concern for school safety, immigration, and other "whole of government" issues. Only three challenges are explored here: the public–private partnership conundrum, information sharing across jurisdictions, global climate change and its impact on natural disasters, and funding of decaying infrastructure.

TABLE 1.6 Some common high- and low-risk hazards are classified according to their consequences[a]

Low-risk hazard	Consequence
S&P500 (1974–1999)	Financial loss
Airline accidents	Deaths
Tornadoes	Deaths
Terrorism	Deaths
Floods	Deaths
Power outage	Megawatts
Earthquakes	Area
Asteroids	Impact area
Pacific hurricanes	Impact area
High-risk hazard	**Consequence**
Hurricanes	Financial loss
Hurricanes	Deaths
Forest fires	Impact area
Small city fires	Financial loss
Earthquakes	Financial loss
Earthquakes	Deaths
Measles	Deaths

[a] Reference [5].

1.8.1 The Public–Private Partnership (PPP) Conundrum

What is the role of the private sector in building resilient systems? What is the responsibility of government during response and recovery? In practice, the public–private partnership (PPP) comes down to regulation and regulatory processes that are determined by politics more than science. For example, the impact of the 1992 EPACT on energy and the electrical power grid, the 1996 Telecommunications Act on communications and the Internet, and the Safe Drinking Water Act (SDWA) of 1974 on environmental regulation profoundly shape the CI sectors, but none of these regulations reduce risk or improve resilience. In some cases, these sectors have become *less resilient* and *riskier* because of regulation.

The National Strategy calls for cooperation between government and private corporations that own and operate much of the most CI systems and KR, but this strategy is at odds with the way government and private companies operate. Government is motivated by politics, while the private sector is motivated by profit. Both parties want security, but they differ in how to achieve it.

Specifically, the 1992 EPACT dramatically weakened the electric power grid by making it unprofitable to improve the transmission assets underlying the grid, and the 1996 Telecommunications Act created the Carrier Hotel architecture that is now recognized as the communications sector's biggest vulnerability. The energy and telecommunications sectors can be improved only through modification or repeal of these regulations, but such radical modifications will require government and the private sector to understand the underlying complexity of these sectors. The necessary expertise does not exist in government and the motivation does not exist in the private sector.

Reversal of deterioration due to aging and wear is a second major factor hinging on PPP. Much infrastructure developed and paid for over the past 120 years is now near the end of its lifecycle. The Interstate Highway System, for example, continues to grow in length as it also crumbles due to inadequate maintenance. The nation's electric power grid is built on 1940s technology and power lines that can no longer support consumer demand. Most drinking water systems in major cities are decaying and slowly failing. Who should pay the mounting maintenance bill?

1.8.2 The Information Sharing Conundrum

Successful infrastructure protection requires information sharing across jurisdictions (*horizontal sharing*) up and down the various tribal, local, state, and federal levels (*vertical sharing*). For example, law enforcement information must freely ebb and flow among and between agencies—local law enforcement must report suspicious activity to regional intelligence centers that report aggregated information to federal agencies. Conversely, situational awareness information and alerts must flow seamlessly from federal agencies to intelligence collection and distribution agencies and finally back to the street level.

Information sharing—both horizontally and vertically—is key to prevention of terrorist attacks and saving lives during a natural disaster. This is why the National Strategy emphasizes, "… protection-related information sharing among private entities within sectors, as well as between government and private entities." These human networks must span tribal, local, state, and federal levels both horizontally and vertically. But information is often hoarded or filtered as if flows in both directions.

1.8.3 Climate Change Conundrum

A third consideration is the rising challenge of global climate change and its impact on CIKR. Clearly the intensity of storms is on the rise, as well as weather-related consequences. The number of billion-dollar natural disasters has outgrown the nation's ability to pay for them, which leads to the question of priorities: "Should we be spending money on target hardening, resilience, and lowering risk when the next super storm is likely to wipe out an entire sector?" Our response to weather and climate change in general may take all of our resources, leaving little to invest in security.

1.8.4 The Funding Conundrum

The national strategy says nothing about how to pay for CIP. And since the private sector exists to make a profit, they are not motivated to invest in target hardening without some financial justification. So what strategy leads to greater security and resiliency through costly enhancements? If we can learn to think asymmetrically about the architecture of infrastructure sectors, why not think asymmetrically about how to finance these needed improvements?

One idea is to "think dual purpose." Can an investment in security serve a dual purpose of also improving ROI? For example, can a private infrastructure sector company reduce operating costs by enhancing security? It might be economically feasible to reduce insurance premiums by decreasing theft at ports. A telecommunications company might increase profits by improving throughput and reliability of telephone calls per hour. Does redundancy in telecommunications also improve the security and reliability of the Internet? Can public schools be converted to hospital rooms during an emergency that requires surge capacity? Can local law enforcement improve service by using online social media and simultaneously reduce the cost of intelligence fusion centers and 911 emergency call centers?

Dual-purpose systems typically achieve greater security through redundancy, because redundancy provides a cushion

against both heavy loading and system failure. Extra standby telecommunications switches and alternate optical fiber lines may seem expensive if not used all the time, but they also provide a high degree of reliability because the system can switch to a backup when needed. Redundant components improve reliability and fill the gap during periods of surge in demand. For example, the New York Stock Exchange was closed for a week following the 9/11 terrorist attacks, because the exchange lacked redundancy. Had the exchange maintained a backup in a separate location, it could have bounced back more quickly.

The funding challenge may actually be an opportunity to rethink infrastructure. Rethinking the power grid in terms of distributed generation and storage reverses the century-old concept of centralized power plants connected to the consumer through an extensive and complex transmission and distribution network. Over the past 40 years, we have learned that the larger the grid is, the harder it falls. Distributed generation can reduce this vulnerability.

1.8.5 Spend 80% on 20% of the Country

The funding conundrum is partially alleviated by realizing that CI is spread unevenly across the country. CIKR assets are concentrated—typically around densely populated areas such as New York City, Silicon Valley, major ports, manufacturing centers, and key rivers and transportation hubs. Moreover, hubs from different sectors are often geographically clustered—typically around a small number of metropolitan areas. For example, Manhattan, New York, has a high concentration of assets in the banking and finance sector. In addition to the New York Stock Exchange, largest Federal Reserve Bank, and many of the world's largest banks, Manhattan is also home to major communication hubs and one-of-a-kind medical centers.

The largest concentration of energy refineries and major source of refined gas and oil products for distribution throughout the United States is located in Galveston Bay, Texas, and along the Louisiana coast. But Texas and Louisiana are also home to the Mississippi supply chain that supplies food and manufactured goods to the rest of the world.

Fairfax County, Virginia, is the home to a large concentration of Internet servers and defense industrial base companies. Chicago is a national hub for transportation and logistics—the sixth largest port in terms of the intermodal supply chain—and also a critical banking and finance center. Most of the 6 million cargo containers that form the backbone of US trade flow through three ports; most of the energy mined to supply fuel for coal-powered power plants is concentrated in Wyoming, and most of the industrial defense base is concentrated in two or three areas of the United States.

These examples suggest an 80–20% rule: 80% of the investment in CIP should be spent on 20% of the country. This, of course, is a political impossibility, but if we are to think asymmetrically about the challenges facing critical infrastructure, we must face reality: target hardening is too expensive to do everywhere. Instead, an optimal strategy invests in the most vulnerable and high-risk parts of the country. If funding is spread equally to all regions of the country, the most critical regions will be under-protected and the other regions will waste the funds.

1.9 EXERCISES

1. What report was the first to use the term "critical infrastructure"?
   a. EO-13010
   b. The "Marsh Report"
   c. The Patriot Act
   d. The National Strategy for Homeland Security
2. How many CIKR sectors and key resources were listed in the Marsh Report?
   a. 5
   b. 8
   c. 13
   d. 18
   e. 16
3. Which agency within DHS did CISA replace in 2018? (Select one)?
   a. NPPD
   b. NIAC
   c. ENIAC
   d. NIPC
   e. PCIPB
4. What sector is not on the list of Table 1.2: CIKR as of 2003 (Select one)?
   a. Agriculture
   b. Internet and the Web
   c. Water
   d. Transportation
   e. US postal and shipping
5. What organization was the first in the United States to advise a US President on critical infrastructure issues (Select one)?
   a. NCS
   b. NSTAC
   c. NIAC
   d. PCCIP
   e. FEMA
6. What federal government agency was the first to be assigned the responsibility of fighting terrorists in the United States?
   a. NCS
   b. NSTAC
   c. NIAC
   d. PCCIP
   e. FEMA

7. When and where was the first bioterror attack on US soil? Who perpetrated it?
   a. 2001: New York City; Al-Qaeda
   b. 1993: New York City; Ramzi Yousef
   c. 1984: Oregon; Ma Anand Sheela
   d. 1995: Oklahoma City; Unabomber
   e. 1995: Oklahoma City; Timothy McVeigh
8. When was critical infrastructure acknowledged as a major component of homeland security? By what document?
   a. 1995: PDD-39
   b. 1996: EO-13010
   c. 1998: PDD-63
   d. 2002: National Strategy for Homeland Security
   e. 2003: National Strategy for the Physical Protection of Critical Infrastructures and Key Assets
9. How many critical infrastructure sectors were defined in PDD-63 in 1998?
   a. 8
   b. 5
   c. 11
   d. 13
   e. 14
10. How many critical infrastructure sectors are defined in the National Strategy for the Physical Protection of Critical Infrastructures and Key Assets in 2003?
    a. 8
    b. 5
    c. 11
    d. 13
    e. 14
11. NIAC was formed in 1999 by EO-13130. What does NIAC mean?
    a. National Industry Advisory Council
    b. National Infrastructure Assurance Council
    c. National Information Assurance Council
    d. National Information Advisory Committee
    e. National Infrastructure Advisory Committee
12. Geographically, critical infrastructure is concentrated around a few locations, which argues for:
    a. Investing to protect dense population centers
    b. Hardening the top 12 metropolitan areas
    c. Investing 80% of the money to protect 20% of the country
    d. Investing most of the money to protect Manhattan
    e. Distribute the generation of power to factories and shopping malls
13. Dual-purpose strategies for coaxing investment in infrastructure protection from the companies that own and operate most infrastructure are defined as:
    a. Enhancing productivity and availability while improving security
    b. Forcing companies to lower insurance policies to pay for improvements
    c. Taxing Internet companies to stop the spread of viruses
    d. Using redundancy to increase volume
    e. Spreading the components of an infrastructure across large geographical areas
14. Hazards can be classified according to their high or low risk according to:
    a. Consequences
    b. Likelihood of disaster
    c. Loss of power and energy
    d. Response versus prevention costs
    e. Emergency response capability
15. The PPP conundrum is:
    a. Companies do not appreciate homeland security.
    b. The private sector is profit driven and government is not.
    c. It is too expensive to protect everything.
    d. CIKR are owned by the private sector, not government.
    e. Companies ignore state and local jurisdictions.

1.10 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. The Department of Homeland Security has an evolving strategy that changes relatively quickly as compared with other governmental agencies such as the National Science Foundation, Department of Defense, and Department of Agriculture. Explain why this is the case and evaluate both pro and con arguments for a shifting strategy.
B. An enduring theme of critical infrastructure protection in the United States has centered on strong leadership from the federal government but with engagement at the state, local, and tribal levels. Alternatives to this vertical integration of governmental control have not emerged beyond early discussions of the National Guard as protector. Is vertical integration the best approach? What are alternatives and why might they provide better security?
C. Immediately following the 9/11 attacks the mantra of homeland security was to protect, defer, respond, and recover. This mantra has disappeared from the discussion over the years leaving most of the emphasis on recovery. Argue either in favor or opposition to this narrowing down of focus. Why isn't protection a bigger piece of the strategy?
D. Qualitative analysis methods are by far more prevalent in critical infrastructure analysis than quantitative methods. The reason is obvious—quantitative analysis is difficult. Argue either in favor of quantitative methods or qualitative methods pointing out pros and cons of each.
E. The Department of Homeland Security employed 225,000 people in 2019 and consumed nearly $50 billion. Is it worth it? What are the alternatives?

REFERENCES

[1] Marsh, R. T. *Critical Foundations: Protecting America's Infrastructures*. The Report of the President's Commission on Critical Infrastructure Protection, October 1997, pp. 3.

[2] U.S. Department of Homeland Security (DHS). National Infrastructure Protection Plan (NIPP): Partnering to Enhance Protection and Resiliency, 2009, pp. 111. Available at http://www.dhs.gov/xlibrary/assets/NIPP_Plan.pdf. Accessed July 29, 2019.

[3] The Whitehouse. *Homeland Security Presidential Directive/Hspd-7*, pp. 1. Available at http://fas.org/irp/offdocs/nspd/hspd-7.html. Accessed December 17, 2003.

[4] Vugrin, E. D., Warren, D. E., Ehlen, M. A., and Camphouse, R. C. A Framework for Assessing the Resilience of Infrastructure and Economic Systems. *Sandia National Labs*, 2010.

[5] Lewis, T. G. *Bak's Sandpile*, 2nd ed, Monterey: Agile Press, 2011.

2

RISK STRATEGIES

Risk analysis is a sophisticated technology developed over the past 250 years to estimate the potential for financial loss in games of chance. In modern times the technology has been applied to a wide variety of disciplines in engineering, social and political science, and, of course, the stock market. At the heart of risk analysis is a simple idea—risk is expected gain or loss under uncertainty. Daniel Bernoulli established *expected utility theory* (EUT) as the earliest known method of quantifying risk in terms of likelihood and gain/loss—R = Pr(C)C, where C is consequence in terms of gain or loss and Pr(C) is the probability of a gain or loss equal to C. Modern risk analysis is descended from Bernoulli's earliest work on EUT. In the field of critical infrastructure protection, probabilistic risk analysis (PRA) was the earliest application of EUT used to assess risk in infrastructure systems. More recently, Bernoulli's EUT has been blending together with Bayesian belief networks, game theory, and probable maximum loss (PML) theory to arrive at today's foundation of risk assessment in homeland security.

The reader in need of a probability primer is advised to review Appendix A, and the advanced reader wanting to understand the mathematical details underlying the survey given here is advised to read Appendix B. The following concepts and results are covered in this chapter:

- *Risk*: Risk analysis is based on EUT, a 250-year old technology invented to predict the results of games of chance. In its simplest form, risk is expected loss: R = Pr(C)C, where Pr(C) is the probability of an event occurring with a gain or loss equal to C. In the study of CIP, C is most often defined as *consequence* from a disastrous event. Consequence can be measured in casualties, dollars, or time. Quantitative risk analysis is a form of *rational actor behavior* that assumes rational people try to maximize their gain or minimize their loss.
- *PRA*: PRA is a simple risk analysis technique originally used in the nuclear power industry to evaluate expected loss due to machinery malfunction, disasters, and even terrorist attacks. In CIP, PRA risk is Pr(C)C or, if the event is an attack by a human, TVC, where T is the probability of an attack, V is the probability the attack is successful, and C is consequence. T is called threat, V is called vulnerability, and C is called consequence regardless of its units of measurement. Because different assets react differently to different threats, a threat–asset pairing is used to identify which threat applies to a particular asset, so that V can be determined. PRA depends on correctly identifying threat–asset pairs and their likelihood of being destroyed with consequence C.
- *FTA*: Fault tree analysis (FTA) is often used to combine threat–asset pairs into a logical structure for risk and vulnerability analysis. A fault tree combines multiple threat–asset pairs into one tree structure using AND, OR, and XOR logic. AND is used to represent redundancy, OR to represent all possible combinations of failure, and XOR to represent single threat–asset pair failures. Fault tree risk minimization produces an optimal allocation of resources to minimize risk. For all

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

practical purposes, faults and failures are two different terms used to describe the same thing—an accident, attack, or natural disaster that causes failure of one or more assets.

- *Kill chain*: In cybersecurity a special form of risk assessment called the kill chain is used to analyze computer security. The idea is to model all paths from the outside of a computer network through levels of software ending up to data accesses that are assumed to be secure by design. These paths form kill chains, which when analyzed in the context of malware have the potential to compromise security. The threat is that malware might successfully penetrate each layer in the chain. The objective is to stop the threat at every level or link along the chain. The kill chain is modeled as a fault tree so that risk calculations and resource allocations can be optimally made.
- *MBRA*: Model-based risk analysis (MBRA) is a software tool for modeling critical infrastructures as fault trees and networks. MBRA calculates risk, computes optimal resource allocation, and simulates single-asset failures and their resulting cascade effects on networks. MBRA is used throughout this book to do the manual labor of calculating risk. However, a spreadsheet can be developed to do the same calculations. MBRA will also be used in subsequent chapters to model infrastructure as a network.
- *ROI*: Making investments in infrastructure is called resource allocation and has a *diminishing return*—the point at which return on investment (ROI) no longer increases as more money is spent to prevent or respond to a CIKR collapse. Diminishing returns must be incorporated into CIP resource allocation, because without it, resources are wasted. MBRA assumes an exponentially declining return as investment increases, but this is a mathematical approximation to reality. ROI is typically measured in terms of reduced risk per invested dollar. The point at which further investment is no longer advantageous is a policy decision, but MBRA can tell the analyst how much each dollar of investment contributes to risk reduction.
- *Limitations of PRA*: PRA is simple and easy to use, but it has many deficiencies. In particular, it assumes threat is always an input to risk. But, in the case of a human threat, it is possible that threat is an output variable. In this case, a rational actor model of threat, vulnerability, and consequence may be used to calculate an optimal attacker allocation to threat T to maximize risk while also calculating an optimal defender allocation to vulnerability V to minimize risk. If PRA is used in this manner, C and budgets are inputs, and R, T, and V are outputs.
- *Game theory approach*: Optimal allocation of attacker resources to maximize risk while also allocating defender resources to minimize risk is a type of two-person game called *Stackelberg competition*. Threat is adjusted by an attacker to increase risk, while vulnerability (and sometimes consequence) is adjusted by the defender to minimize risk. Once an equilibrium between attacker and defender is reached, threat T and vulnerability V are both output values. There is no need to know T and V in advance. Game theory techniques are used to optimally allocate limited investments (resource allocation) to minimize risk on the part of the defender and maximize risk on the part of an attacker.
- *Conditional probability analysis*: *Bayesian network (BN)* analysis is a second method of obtaining threat T rather than assuming it is an input to PRA. A BN is a network of beliefs—expressed as *propositions*—that are **true**, **false**, or something in between. BN analysis uses conditional probabilities and mounting evidence (expressed as a probability of an even happening or not) to arrive at threat T. BN improve on the accuracy of a prediction as more evidence is gathered and plugged into the BN—a model of reality.
- *Exceedence*: The PRA model is less useful for representing risk from natural disasters, where humans are victims and not actors. T and V have little meaning when analyzing the risk of damage due to a hurricane or earthquake. Therefore, another approach based on exceedence probability is preferred. Exceedence probability EP($x \geq$X) is the probability that x equals or exceeds a certain value, X. Ranked exceedence EP($n \geq$N) represents the probability that n events equal or exceed N. True exceedence EP($c \geq$C) represents the probability that the size of an event c equals or exceeds C. Ranked exceedence is used to count likely events, while true exceedence is used to estimate the likelihood of an event. Both methods are used to estimate likelihood in the equation for PML.
- *Ranked exceedence*: The Gutenberg–Richter law for earthquake magnitude is an example of a ranked exceedence probability, because it relates the number of earthquakes of size M or larger to their frequency. For example, M = 5.6 equates with the *number of earthquakes* of a certain size—not their size. However, the size of an earthquake can be computed from the Gutenberg–Richter scale if you know M. Thus, ranked exceedence is the probability that an event equals or exceeds a certain rank order, among all known events.
- *True exceedence*: Large flood exceedence probability is an example of a true exceedence probability, because it relates the probability of a single flood equal to or greater than size C, where C is typically measured by the discharge (cubic meters per second). For example,

the probability of a 100-year flood (or greater) is given by the true exceedence probability. If a 100-year flood occurs every 30 years, its true exceedence probability lies on the 30-year mark of the x-axis and is equal to the corresponding y-axis, for example, 1% or whatever is known historically.

- *Power laws*: In most cases studied in this book, exceedence probability curves obey a long-tailed power law of the form $EP \sim x^{-q}$, where q is the *fractal dimension* of EP and x is a consequence, distance, or elapsed time between catastrophic events. The exponent q is also a proxy for resilience, because systems subject to collapse according to EP are more resilient if q is large and more fragile if q is small. Thus, fractal dimension relates indirectly to the likelihood of an event of a certain size occurring.
- *PML risk*: PML is a more fitting definition of risk for natural disasters because natural disasters are accurately modeled by exceedence probabilities. There is no T or V value associated with natural disasters. PML risk depends on EP(C) rather than TV and is defined as the product $EP(c \geq C)C$. And since $EP(c \geq C)$ is a power law, PML risk is a function of consequence C and fractal dimension q. PML $R = C^{(1-q)}$. Note that boundedness of PML R is determined by q: if q is less than 1, PML R is unbounded. If q is greater than one, PML R diminishes to 0.
- *Fractal dimension*: *Black swan events* are disastrous incidents that rarely occur but have large consequences. They are high-consequence, low-probability events. What is PML risk when EP is very near zero and C is very large? When fractal dimension of PML risk is less than one (fragile), black swan risk becomes unbounded so that larger risk is associated with larger consequences. When fractal dimension is greater than one (resilient), black swan risk approaches zero as consequences increase. Therefore, the fractal dimension of high-risk hazards is $q < 1$ and of low-risk hazards is $q > 1$. Fractal dimension q defines high- and low-risk hazards.
- *High and low risk*: Hazards can be classified as either high risk or low risk, depending on their fractal dimension. The largest known floods over the past 1.8 billion years are high risk because $q < 1$. Terrorist attacks by al Qaeda during the period 1993–2009 are low risk, because $q > 1$. The global pandemic severe acute respiratory syndrome (SARS) was low risk, but cyber exploits are high risk. Fractal dimension defines the boundary between low and high risk.
- *Risk strategy*: The optimal CIP risk strategy attempts to reduce risk by reducing threat, vulnerability, and consequence and, more importantly, by increasing fractal dimension q. Because of diminishing returns, the best risk strategy manages a CIKR portfolio of assets and systems and spreads resources across multiple assets and systems.

2.1 EXPECTED UTILITY THEORY

Daniel Bernoulli (1700–1782)—a third-generation grandson of the famous family of Swiss mathematicians—laid the foundation of modern risk analysis when he formulated *EUT* in 1738. According to Bernoulli, risk is the product of the probability of a certain outcome and its consequence: R = Pr(C)C, where Pr(C) is the probability of losing C dollars, say, and C is the loss measured in dollars. When n independent events are possible, risk is simply the sum of all expected values of R. This breakthrough in risk calculation continues to be used today in financial and engineering calculations.

Of course consequence can also be measured in terms of fatalities, economic decline, loss of productivity, and other measures. Probability can be calculated in a number of ways (discussed in detail in Appendix A and B), but it is always a unit-less number in the interval [0, 1]. Thus, risk is measured in the same units as consequence. If consequence is given in terms of fatalities, then risk is given in terms of loss of life. If measured in terms of dollars, then risk is expressed in terms of dollars. Risk is typically measured in dollars, here, because human lives, time, and so on can be converted into dollars.

It is important to note that risk is not a probability and probability is not a risk. Rather, the elements of risk are likelihood as measured by a probability and gain/loss as measured by a consequence. For example, the likelihood of having a computer virus attack your personal computer is rather high, but the risk is rather low if we measure consequence as the cost associated with removing the virus. On the other hand, the likelihood of another 9/11-sized terrorist attack is rather small, but the consequence is very high. The risk of any event is large if the product of likelihood and consequence is large but small if the product is small. Probability and consequence are handmaidens in the estimation of risk—both are needed to calculate risk.

Risk is also not vulnerability or threat. These two terms are often mistaken for risk, because they are closely related to risk. Generally, vulnerability is a weakness in an asset that may be exploited to cause damage. It can be quantified as a probability, but it is not risk, because it is not expected gain or loss. Similarly, threat is a potential to do harm that can also be quantified as a probability, but it is not a form of risk for the same reasons as vulnerability. Generally, threat can be quantified as the probability of an attack or catastrophic event and assigned a number between zero and one. But as discussed later, this definition of threat is controversial.

To be clear, threat and vulnerability are probabilities, here, and consequence is a form of damage measured in any number of different ways. To obtain an estimate of risk, we must multiply threat, vulnerability, and consequence together. This is one of several ways to obtain risk, but not the only way.

Threat is typically associated with human attacks—terrorism—while natural disasters are typically associated with a hazard such as an earthquake or hurricane. When convenient, threat and hazard will be used interchangeably, here. In both cases, an asset must be associated with threat or hazard to make sense. Thus, threat–asset pairs such as malware hacker–Internet Web site, hurricane–house, thief–bank, and so on must be paired together before Pr(C), T, V, or C has any meaning.

We are now in a position to understand the modern manifestations of risk-informed decision-making used to decide how best to allocate resources to reduce expected losses. The challenge of risk assessment comes down to the challenge of calculating probabilities and consequences. How do we estimate the probability of a future event, and how do we know the extent of damages? As it turns out, this is more complicated than the pioneers of EUT ever imagined.

2.1.1 Threat–Asset Pairs

The most fundamental unit of risk can be found in the *threat–asset pair* as illustrated in Figure 2.1. As an example, suppose the asset is your car and the threat is a nail in the road that leads to a flat tire. This threat–asset pair is represented in Figure 2.1a as two blocks with a line connecting them. For each threat–asset pair, we can estimate the probability of the threat occurring—the hazard in this case, because it is a nonterrorism event—and the probability that the asset will fail, given that the hazard occurs. In addition, we can estimate the consequence of failure. Risk is the product of these two properties of the threat–asset pair.

In this example, the probability of puncturing the tire with a sharp object in the road is converted to a number, T, and the probability that the tire deflates if punctured is vulnerability, V, and the consequence, C, is the damage caused by the flat tire. If we know the likelihood of T and V and the value of damages C, we can multiply them together to obtain risk, R.

According to the DHS glossary:

> THREAT: natural or man-made occurrence, individual, entity, or action that has or indicates the potential to harm life, information, operations, the environment and/or property.

This definition is quantified by assigning a probability to the natural or man-made occurrence:

> THREAT, T: probability of a natural or man-made occurrence or action that has the potential to harm life, information, operations, the environment and/or property.

Similarly, the layman's definitions of V and C are modified to quantify them so they can be used to calculate risk:

> VULNERABILITY: physical feature or operational attribute that renders an asset likely to fail due to a given hazard—the probability of failure if attacked or subjected to the threat.
> CONSEQUENCE: effect of an event, incident, or occurrence—damages due to a failure—typically measured in dollars, casualties, or lost time.

These modifications to the layman definitions allow us to compute risk and quantify the expected loss due to an event caused by a threat applied to a specific asset. That is, risk is the product of T, V, and C and is measured in the same units as C. Table 2.1a shows results of applying simple TVC to numbers supplied by the example in Figure 2.1.

2.2 PRA AND FAULT TREES

The father of modern risk assessment as it applies to homeland security was an MIT professor of nuclear engineering, Norman Rasmussen (1927–2003). He achieved notoriety in the 1970s by debating the safety (or lack of it) of nuclear power plants with Ralph Nader. The televised debate took place in 1976, 3 years before the Three Mile Island (TMI) nuclear power plant meltdown.

Perhaps more important than the debate with Nader was the method of risk assessment employed by Rasmussen—now known as *probabilistic risk analysis*. His 1975 report defined risk as the expected loss due to a failure: risk = Pr(failure)C(failure), where Pr(failure) is the likelihood of a reactor failing and C(failure) is its consequence. Rasmussen's use of EUT in PRA is easy to understand, but it can be difficult to apply in practice, especially if power plant operators cannot calculate Pr(failure) and C(failure). Where do Pr(failure) and C(failure) come from?

For example, the TMI nuclear power plant meltdown was supposed to be impossible. Thus, Pr(failure) was supposed to be zero. In hindsight, Pr(failure) is not zero, but how does an operator know this beforehand? If we use a priori analysis, we must know all of the ways failure can happen and all the ways it cannot. If we have historical data to support an a posteriori estimate based on the major nuclear power plant catastrophes that have occurred over the past 60 years, we can substitute histogram data from historical observations into Rasmussen's PRA formulation to get risk. (This is left as an exercise for the reader.[1])

Estimating consequences is somewhat easier, but not straightforward. The TMI meltdown caused approximately $2.4 billion in property damage and $1 billion in cleanup costs. Although a number of studies were

[1]http://en.wikipedia.org/wiki/List_of_civilian_nuclear_accidents

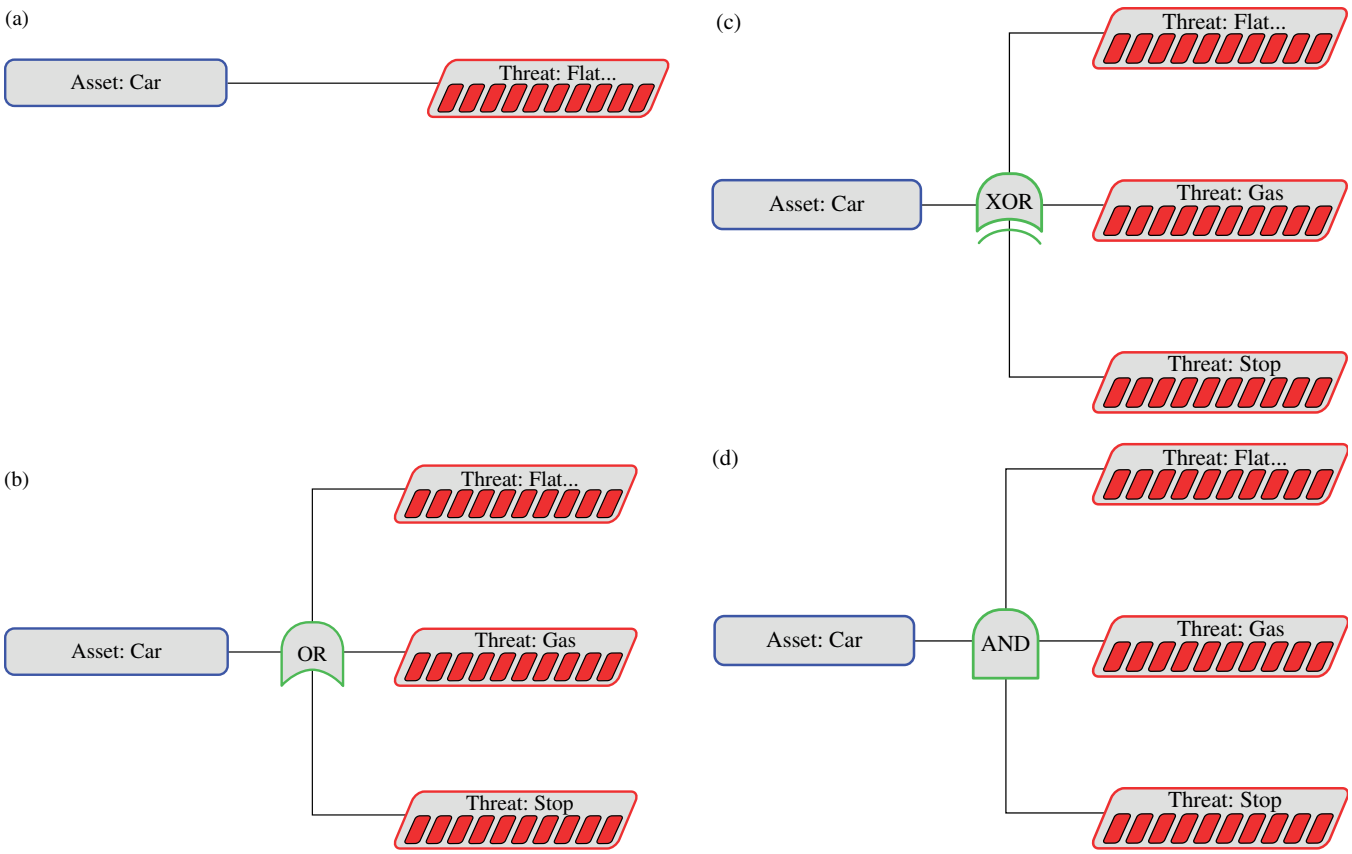


FIGURE 2.1 The fundamental unit of risk is the threat–asset pair as illustrated here. (a) Basic threat–asset pair: car is the asset; flat tire is the threat. (b) Three threat–asset pairs: flat tire, empty gas tank, and stalled engine. The three pairs are connected by an OR gate representing the possible occurrence of 0, 1, 2, or all 3 hazards simultaneously. (Note: there are eight possible combinations that stop the car.) (c) The three pairs are connected by an XOR gate representing the possible occurrence of only one hazard at a time: either flat tire, empty gas tank, or stalled engine. (Note: there are three possible combinations that stop the car.) (d) The three pairs are connected by an AND gate representing the possible occurrence of all three hazards simultaneously. (Note: there is only one possible combination that stops the car.)

TABLE 2.1 The parameters for the threat–asset pairs in Figure 2.1 include T, V, and C as well as the costs to eliminate risk by reducing V

(a) Initial values of T, V, C, and elimination cost yield a total risk of $270.00. Fault tree vulnerability is 66.25%

Threat	T (%)	V (%)	C	Elimination cost	Initial risk
Flat	50	50	$300	$100.00	$75.00
Gas	80	50	$300	$50.00	$120.00
Stop	25	100	$300	$200.00	$75.00
				Total risk	$270.00

(b) Allocation of $50 minimizes risk to $117.83 by optimally reducing V. Fault tree vulnerability is 34.8%

Threat	T (%)	Reduced V (%)	Elimination cost	Allocation	Reduced risk
Flat	50	24.55	$100	$18.18	$36.83
Gas	80	7.67	$50	$23.96	$18.42
Stop	25	83.43	$200	$7.87	$62.58
				Total risk	$117.83

conducted to assess the health consequences on the people living in the region, the consequences of radiation exposure have never been fully quantified. Furthermore, the nuclear power industry suffered for decades following the incident. How does one put a dollar value on "loss of business"? This illustrates the difficulty of estimating consequences, even when the risk method is as simple as Rasmussen's PRA.

Rasmussen addressed the problem of estimating Pr(failure) using an old engineering technique called *fault tree analysis*. Instead of attempting to calculate Pr(failure), directly, he decomposed the components of each power plant into simple threat–asset pairs and then inserted them into a fault tree to determine the likelihood of the entire power plant failing if any one of its components failed. Rasmussen defined Pr(failure) for each threat–asset pair, i, as the product of threat and vulnerability, $t_i v_i$, and risk contribution of each threat–asset pair as the product, $t_i v_i c_i$.

PRA also involves a *logic model* of how threat–asset pairs are combined together to get total risk over all threat–asset pairs. A *fault tree* is a set of threat–asset pairs combined together using AND, OR, and XOR logic. The leaves of a fault tree are threats, and the nodes are assets or leaves to a higher-order tree. Leaves are connected via logic gates representing one or more possible threats. Consider the following application of FTA to the car-and-tire example.

2.2.1 An Example: Your Car

To make the threat–car example more interesting, suppose your car is subject to three hazards—a flat tire, empty gasoline tank, and engine failure. These three threat–asset pairs are combined together in a logic-based model of how your car might fail. In Figure 2.1b, an OR gate is used to combine the threat–asset pairs. This OR fault tree represents how your car might fail because of zero, one, or any combination of the three threats occurring individually or in combination. The logic gate OR means "zero or more combinations" of threats.

Similarly, an XOR fault tree is constructed with an XOR-gate connector, representing single failures (see Fig. 2.1c). This means only one of the three hazards is able to stop your car. Interestingly, the XOR fault tree represents lower risk, because it excludes the occurrence of two of the three threats for each possible failure mode. That is, the probability of a flat tire *excludes* the probability of running out of gas and engine failure. Similarly, the possibility of an empty gasoline tank excludes a flat tire and engine failure, and the possibility of an engine failure excludes the possibility of the other two threats. The X in XOR means *exclusive*. As a result of the exclusivity of single hazards, XOR fault tree risk will be lower than the OR tree risk.

Figure 2.1d illustrates a third form of fault tree: the AND tree. In order for your car to fail in this logic model, all three hazards must occur—a flat tire, an empty gasoline tank, and an engine failure. If any one of the three threats is absent, your car may be injured but it will not fail. The AND fault tree represents *redundancy*, because all three threat–asset pairs must fail; otherwise one or two component failures avoid car failure. Your car will operate with a flat tire, or an empty gasoline tank, or an engine failure, or any two of the three threats. (Note: you can still push it, even if it is out of gasoline or the engine fails!)

How risk is calculated from these fault tree models is explained in mathematical detail in Appendix B. Figure 2.1 illustrates how the fundamental threat–asset pairs are combined into a fault tree for each of the possible models described above. For example, Figure 2.1b illustrates how to model one or more hazards using the OR gate. The probability that your car will fail due to one or more of the threats occurring is 66.25%. Table 2.1 illustrates how risk is computed. Alternatively, the reader can download MBRA software to perform these calculations automatically.

2.3 MRBA AND RESOURCE ALLOCATION

Tools like MBRA exist to perform risk and fault tree vulnerability calculations.[2] In addition, MBRA includes algorithms to compute the optimal allocation of a fixed budget to minimize risk. A user inputs T and V probabilities, consequences, and vulnerability elimination costs, and MBRA returns an optimal investment strategy for minimizing risk. The optimal allocation reduces vulnerability V but leaves T and C unchanged. (The investment is applied to reduce V.)

Figure 2.1b illustrates how MBRA allocates $50 to minimize the likelihood that your car will fail because of the tree hazards. The same T, V, and C values as before are used, but now the *cost* of avoiding or preventing each threat is incorporated. The elimination cost is an estimate of how much it cost to prevent each threat–asset pair from occurring by reducing V. For example, purchasing a spare tire for $100 eliminates the flat tire hazard; the empty gasoline tank hazard is eliminated by purchasing a backup can of gasoline to put in the trunk for $50; and the engine failure hazard is prevented by scheduling maintenance at a cost of $200.

Table 2.1b shows the results of vulnerability *buy down*. A budget of $50 is apportioned across all three threat–asset pairs to reduce risk by reducing V. Assuming an OR fault tree and $50 budget, a minimum risk of $117.83 is obtained by investing $18.18 to eliminate the flat tire hazard; $23.96 to eliminate the empty gasoline hazard; and $7.87 on engine failure prevention. Of course, these are partial eliminations,

[2]MBRA downloads are at: www.CHDS.us/resources.

because a total of $350 would be needed to completely eliminate all hazards.

Fault tree vulnerability—the probability of at least one threat occurring and thereby stopping your car—declines from 66.3 to 34.8% after investing $50 and declines even further to 11.2% after investing $150. How much investment is enough, and when does diminishing returns render more investment a waste of money? This is answered by evaluating ROI.

Fault tree risk is reduced from $270 to approximately $117, for an ROI of ($270–117)/$50, or $3.06/$. Generally, more budget means more risk reduction, but risk reduction has a *diminishing returns* (see Fig. 2.2). The MBRA fault tree model assumes an exponential decline in vulnerability as investment increases. And, because an infinite investment is required to reduce an exponential to zero, some vulnerability will always remain.

The first dollar invested has a higher ROI than the last dollar. This shows up in Figure 2.2 as an exponentially declining risk versus total budget, and an ROI curve that also declines exponentially. An investment of $100 yields an ROI of approximately $2.00/$. Therefore, the amount of investment, and the corresponding amount of ROI achieved, is a policy decision. Figure 2.2 contains a vertical dotted line at approximately $100 corresponding with a policy decision to get $2.00 in risk reduction for each $1.00 invested. Therefore, an ROI policy might trade improvement in security for cost—the more you pay the more you get. But ROI also declines with risk, so ROI is itself a diminishing return asset.

2.3.1 Another Example: Redundant Power

Consider a more elaborate and realistic example. Suppose a hospital wants to evaluate the utility of backup power as a way to increase resiliency through redundancy. During normal operation, the hospital runs on electrical power provided by the grid. But when the grid fails, the hospital switches to its backup system—a diesel-powered generator. The hospital cannot fail unless both sources of power fail. The backup generator is a redundant resource, and therefore the fault tree contains an AND gate as shown in Figure 2.3.

The AND fault tree for this hypothetical hospital has three levels—the hospital, component level containing backup and grid, and the threat level containing flood, terrorist, and mishap. In general, fault trees can have many levels of components, which allows the analyst to model complicated systems. In this example, the AND logic composes backup and grid. In turn, the backup component contains one threat–asset pair: flood–backup. The grid component combines two threat–asset pairs: terrorist–grid and mishap–grid. The two grid pairs are combined using the OR logic shown in Figure 2.3.

Table 2.2a shows the inputs for each threat–asset pair in Figure 2.3. Initial estimates of T and V are 50%, which represents maximum uncertainty. That is, there is not enough information to know whether T and V should be high or low. Therefore, the maximum ignorance values are used. Consequences are obtained by calculating the financial damages to the hospital due to each hazard—flood, terrorist attack, and mishap. Elimination costs are calculated based on the cost of vulnerability reduction as before.

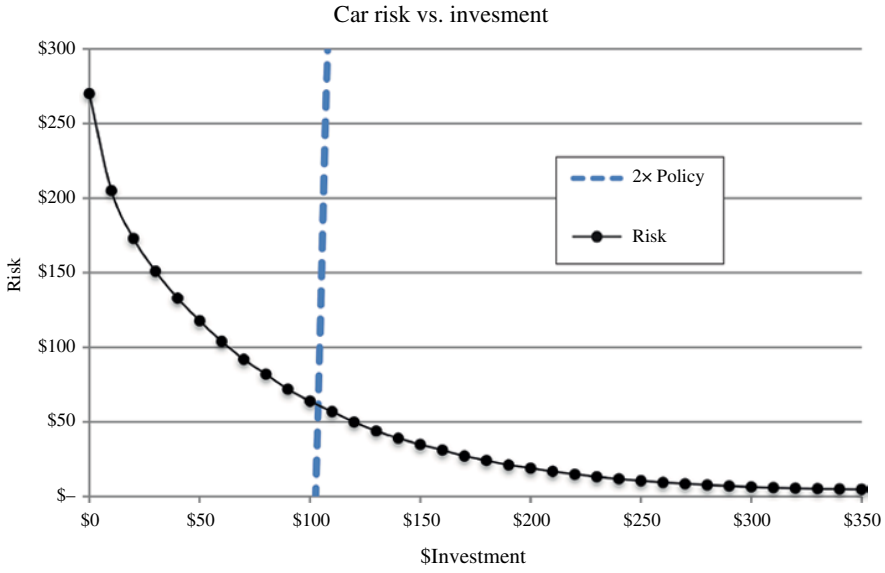


FIGURE 2.2 Risk and return on investment decline after a modest investment in vulnerability reduction. This graph was obtained for the OR fault tree resource allocation algorithm in MBRA that assumes an exponential diminishing returns relationship between budget and vulnerability. The vertical dotted line shows that $100 invested returns $200 in risk reduction for an ROI of $2/$.

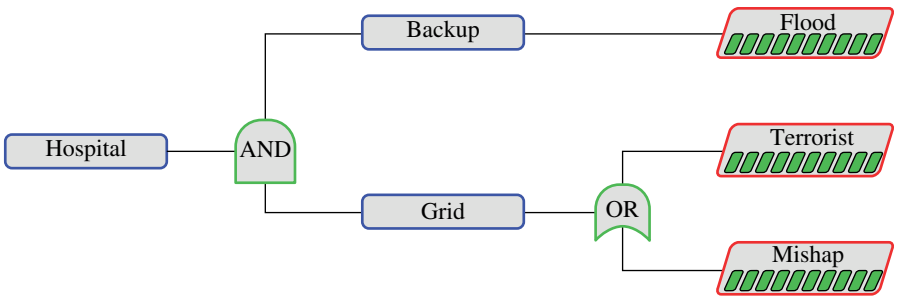


FIGURE 2.3 AND fault tree for the hypothetical hospital power supply: redundancy is modeled by placing an AND gate between the two sources.

TABLE 2.2 Inputs and analysis of the hospital fault tree in Figure 2.3 shows investing $125 thousand reduces risk from $1000 thousand to $581.5 thousand

(a) Initial inputs and risk. Dollars are in thousands

Threat	T (%)	V (%)	C	Elimination cost
Flood	50	50	$1000	$100
Terrorist	50	50	$2000	$1000
Mishap	50	50	$1000	$200
Initial risk	$1000			

(b) Results of investment of $125 thousand to secure the hospital's power. Initial risk of $1000 thousand is reduced to $581.5 thousand

Investment: $125 Threat	Allocation	Reduced V (%)	Reduced risk
Flood	$44.00	8.95	$44.73
Terrorist	$28.50	44.70	$447.30
Mishap	$52.55	17.90	$89.46
Reduced risk	$581.49		

The numbers in Table 2.2b show the results of investing $125 thousand to reduce V and, in turn, risk from $1000 thousand to $581.5 thousand—a 41.85% reduction. The ROI is $3.35/$. But, if the policy is to obtain an ROI of $2.00/$, as shown in Figure 2.4, $350 thousand would have to be invested, reducing risk from $1000 thousand to $295 thousand.

Hospital failure probability is initially 10.94%, because both grid and backup must fail in order to do harm to the hospital. An investment of $125 thousand reduces it to V = 1.3%, and an investment of $350 thousand reduces V to 0.35%. That is, vulnerability is very low because of the redundancy of backup power.

This example illustrates the value of redundancy. Without the redundant backup, risk due to grid failure would be $750 thousand. If the backup system were combined with the grid using an OR model, the risk due to one or the other failing would still be $581.5 thousand, assuming a vulnerability reduction investment of $125 thousand as before, but the likelihood of one or the other failing would be 32.5%, with the OR fault tree, instead of 10.94% with the AND tree. Redundancy is three times more effective in reducing vulnerability of the hospital than reliance on the grid, alone. AND redundancy is very powerful medicine for this hospital.

The two examples—car and hospital—illustrate contrasting models of risk reduction. In both examples, risk declines exponentially because MBRA optimization assumes exponential diminishing returns. (This is an artifact of MBRA, but is it true in reality?) But the decline in vulnerability is much sharper for the redundant hospital power fault tree than the car fault tree. This is due to the redundant AND gate in the case of the hospital.

2.4 CYBER KILL CHAINS ARE FAULT TREES

The *cyber kill chain model* aims to organize threats and risks associated with attacks on computer systems so they may be addressed in a systematic and structured manner. Threats that go far beyond script kiddies' exploits and amateur hacks are called *advance persistent threats* (APTs). The common objective of an APT is to insert a remote access Trojan (RAT)

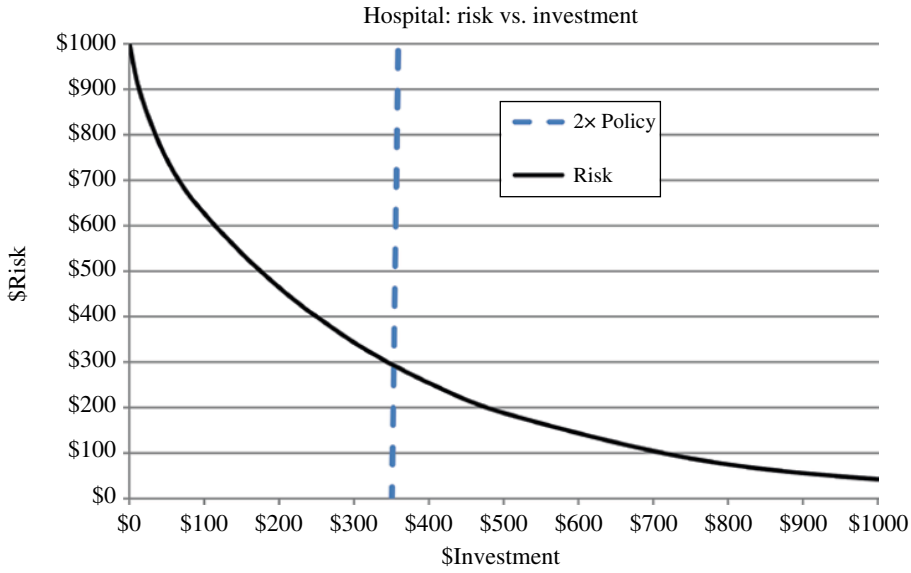


FIGURE 2.4 Return on investment analysis for the redundant hospital power model in Figure 2.3. The vertical dotted line shows where an ROI of $2.00/$ is achieved.

into a victim's computer system so that the adversary can take control anytime she wishes. Thus, kill chains can be modeled as special fault trees because threats are APTs, computer weaknesses are vulnerabilities, and exploits have consequences.

Hutchins *et al*. [1] define a cyber kill chain as a sequence of intrusions leading up to destructive action on the part of an APT:

1. *Reconnaissance*—Research, identification, and selection of targets, often masquerading as benign Internet Web sites such as conferences and mailing lists for email addresses, social relationships, or information on specific technologies.
2. *Weaponization*—Packaging an RAT and exploit into a deliverable payload, typically by means of an automated tool (weaponizer). For example, client application data files in Adobe Portable Document Format (PDF) or Microsoft Office document format serve as the weaponized deliverable.
3. *Delivery*—Transmission of the weapon to the targeted network or machine. For example, in 2004–2010, the three most prevalent delivery vectors for weaponized payloads by APT actors were email attachments, Web sites, and USB thumb drive removable media.
4. *Exploitation*—The adversary triggers the payload code after it is delivered to the victim host. Frequently, exploitation targets an application or operating system vulnerability, but it could also more simply exploit the users themselves or leverage an operating system feature that automatically executes the intruder's code.
5. *Installation/spread*—Installation of an RAT or backdoor on the victim's system allows the adversary to maintain persistence inside the environment. The RAT is controlled and activated by the adversary at any time.
6. *Command and control (C2)*—Typically, APT malware requires manual interaction rather than automatic control. Intruders have "hands on the keyboard" access inside the target environment.
7. *Actions on objectives*—Typically, the objective is data *exfiltration*, which involves collecting, encrypting, and extracting information from the victim environment, as well as violations of data integrity. Alternatively, the intruders may only want to hop to another system or move laterally inside the network.

Note that all seven steps must be completed before the exploit is successful. In terms of FTA, these steps are connected by an AND gate (see Fig. 2.5). A typical multi-path kill chain is shown in Figure 2.6, assuming a kill chain is established between a system and three different access points: a user inserting a USB thumb drive into the system, a third-party vendor with password access, and a typical user that may be a victim of spear phishing. A kill chain is established for every trusted path connecting users with their data.

2.5 PRA IN THE SUPPLY CHAIN

One of the most successful applications of PRA outside of the nuclear power industry is MSRAM (*Maritime Security Risk Assessment Method*)—a US Coast Guard method and tool for assessing port security [2]. MSRAM incorporates tools for estimating T, V, and C utilizing a modified PRA model:

$$\text{Risk} = \text{TVC}$$

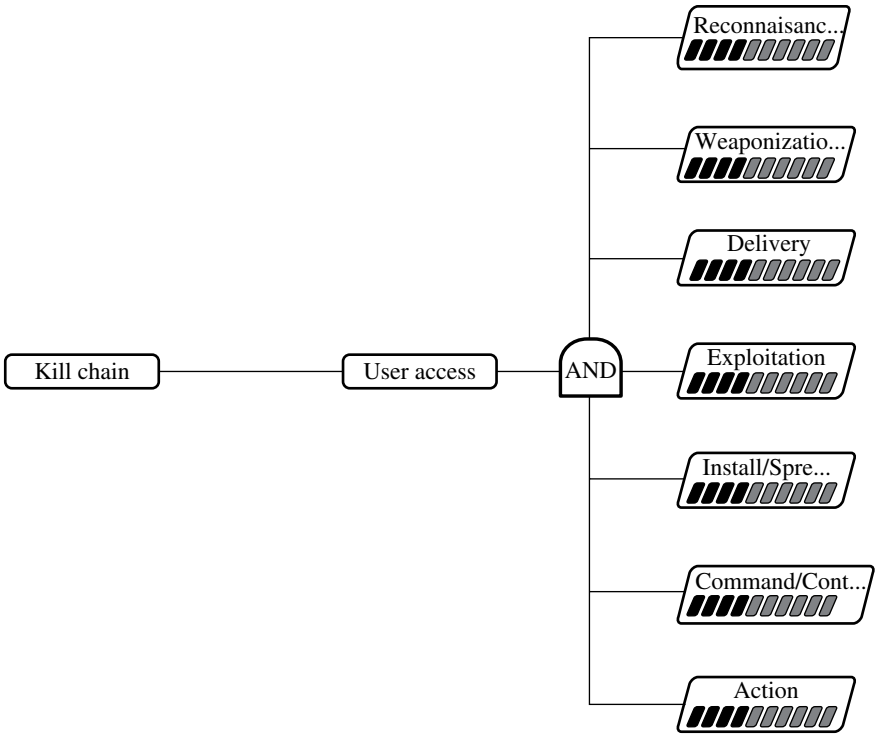


FIGURE 2.5 A fault tree model of a single trusted path from a user to data contains kill chain steps connected by an AND gate. All steps must be successful for an intruder to gain control.

where T is INTENT × CAPABILITY

INTENT is a measure of propensity to attack

CAPABILITY is a measure of ability to successfully attack

V is a measure of target weakness

C is modified consequences, moderated by preventive measures

In MSRAM, T is a combination of a terrorist's *intent* and *capability* to carry out an attack. V is a measure of vulnerability due to a lack of prevention, lack of target hardening and mitigation, and lack of resiliency. Consequence, C, is actually a reduced consequence calculated by considering how well port authorities collaborate with sister law enforcement agencies, the port's response capability, and other target-hardening factors. T, V, and C are obtained by selecting options from a list, so the user does not have to enter numbers.

MSRAM is scenario driven, meaning it takes a user through a scenario as it collects inputs. One scenario might describe a plot to block a port by ramming and capsizing a large ship. Another scenario might describe an IED attack on key assets. MSRAM supported approximately a dozen scenarios at the time this was written. Scenarios simplify the task of estimating T, V, and C and keep users focused on incidents of interest to the Coast Guard.

MSRAM produces a risk index number (RIN) obtained from multiplying the MSRAM scenario-driven estimates of T, V, and C. It is also part of more elaborate risk assessment tools used by the USCG to allocate resources such as people and ships. For example, PROTECT uses MSRAM RIN numbers to randomize and schedule patrols to fend off poachers, terrorists, drug runners, and smugglers. PROTECT is a game-theoretic tool described in more detail in Chapter 16.

MSRAM has been used to analyze thousands of assets in ports across the county. RIN values are collected at the local port level, regional Coast Guard levels, and USCG headquarters. By comparing similar assets across all ports, analysts can standardize the results—an important feature that allows headquarters to compare RINs across the country. The RIN values are ranked from highest to lowest to determine resource allocation. While this is known to be nonoptimal, it does reduce the highest risks before allocating limited resources to lower-risk assets.

2.6 PROTECTION VERSUS RESPONSE

PRA, MSRAM, and FTA methodologies have limitations. First, there is no consideration of the cost involved in reducing risk by reducing vulnerability or consequence as in

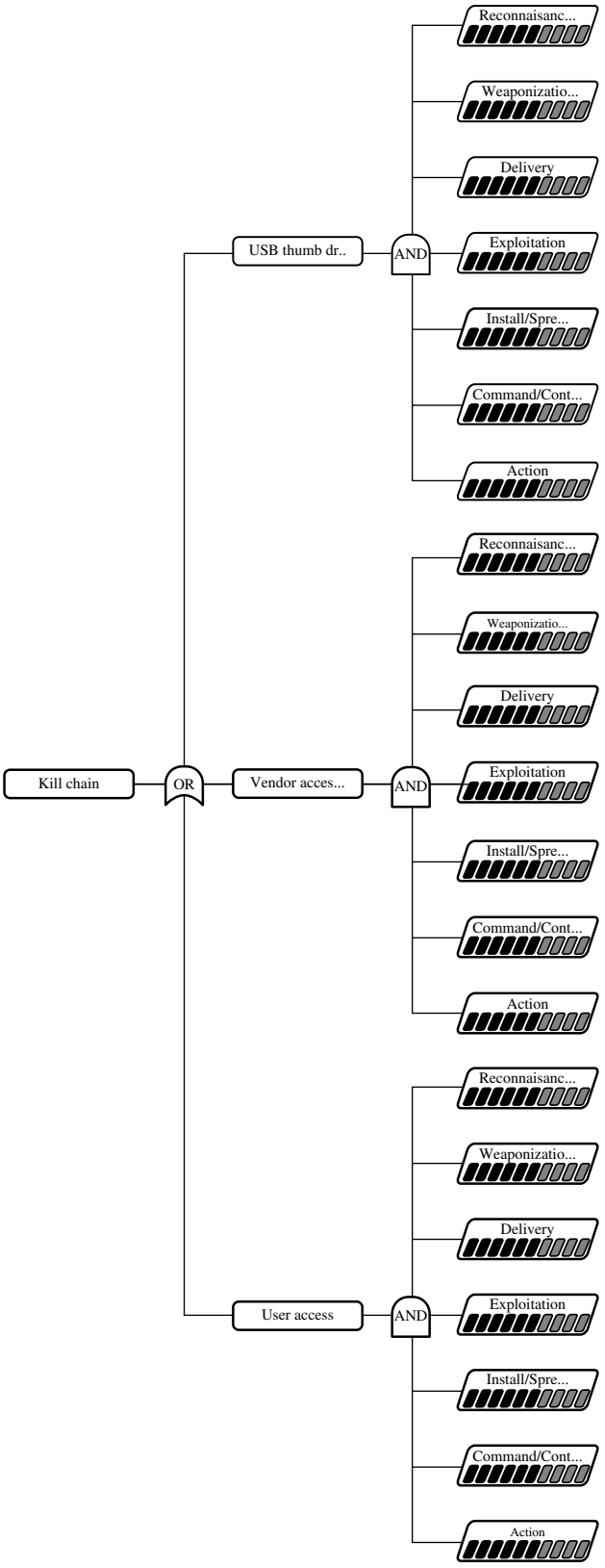


FIGURE 2.6 A multipath kill chain connects paths with an OR gate and replicates the seven steps of the kill chain for each path. Any one or multiple path can be compromised, resulting in a successful exploit.

the MBRA fault tree model. MSRAM has no elimination or mitigation cost capability at all. (MBRA applies investments to vulnerability reduction only, which is an incomplete model, too.) Lacking any method or guidance for optimal resource allocation, most operators simply rank assets according to risk and then apply resources to the highest-risk-ranked assets. This is famously nonoptimal in general, because different threat–asset pairs cost different amounts to protect and respond to.

For example, the cost to harden the Golden Gate Bridge is much higher than the cost to harden a Google Internet server. Even if the risk to both Golden Gate Bridge and Google Internet server were identical, it is more rational to allocate more resources to an asset with lower elimination cost, because overall risk reduction is greater when investing in the less expensive asset. Higher ROI is typically associated with lower prevention and response costs, which favors investment in higher ROI assets, regardless of their contribution to risk. Of course, there is great political and esthetic value attached to the Golden Gate Bridge, which is not easily quantified and entered into a calculator.

One remedy to this imbalance is to incorporate prevention, vulnerability, and response costs in the expected utility definition of risk. Investment in prevention might lower threat; investment in vulnerability might increase resilience; and investment in response might lower consequence. Investments are applied separately to prevention, resilience, and response, respectively. Risk is reduced by a combination of vulnerability and consequence reduction. (A similar argument can be made for reducing threat as well.)

This level of detailed investment has been implemented in commercially available tools such as *NetResilience*.[3] In fact commercially available tools often breakdown T, V, and C into more detailed models of each. MBRA's network model—briefly introduced here—divides elimination cost into resilience and response costs. Then, resilience investments are used to reduce V, and response investments used to reduce C. For example, a network model of the redundant power source for the hospital in Figure 2.3 is shown in Figure 2.7. In place of a fault tree, each component—grid and backup—is represented as a node with five properties: t, v, c, prevention cost pc_i, and response cost rc_i. Table 2.3 contains example input values and the results of resource allocation such that V and C are reduced in order to minimize risk. Threat, T, remains unchanged—a topic addressed later.

Table 2.3 summarizes the input values and calculations obtained from MBRA's resource optimization algorithm. MBRA sequentially optimizes to obtain minimum risk.

[3]www.critsci.com

First, the prevention budget is applied to reduce vulnerability, followed by application of the response budget to reduce consequence. Both reductions assume the same exponential diminishing returns curve used by MBRA's fault tree algorithm. This may introduce inaccuracies when the rate of consequence reduction differs from the rate of vulnerability reduction.

As before, expected utility is the sum of risks—$1000 thousand before investments in risk reduction and $446.8 thousand after optimization. This reduction was obtained by investing $75 thousand in prevention and $50 thousand in response. This yields an ROI of $4.43/$. Why is this much higher than the previous ROI of $3.35/$? The answer is that consequence reduction contributes more to risk reduction than does vulnerability reduction in this example.

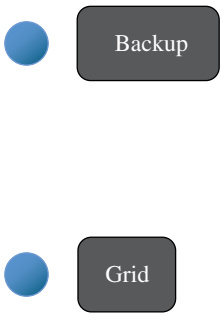


FIGURE 2.7 MBRA's network model of the hospital redundant power source requires only two nodes: backup and grid.

2.7 THREAT IS AN OUTPUT

A second major criticism of PRA concerns the placement of T on the right-hand side of the risk equation, R = TVC. Critics say that T should be an *output* rather than an input to risk assessment. That is, threat should be a function of vulnerability, because terrorists are more likely to attack weaker targets than stronger or better-protected targets. According to the critics of PRA, a rational terrorist will attack the most vulnerable target to maximize his or her expected utility.

But what if an intelligent adversary looking to maximize risk in a competition with a defender attempting to minimize risk adjusts threat to take advantage of the defender's weaknesses? This formulation sets up a two-party competitive game. One party is the defender, and the other party is the attacker. The defender attempts to minimize risk by reducing V and C. The attacker attempts to maximize risk by increasing T. When this game is applied to the hospital grid and backup system, allocations of both defender and attacker budgets must accommodate the competition as well as diminishing returns. Such a competition is called *a Stackelberg game* and leads

TABLE 2.3 Inputs and resource allocation calculations minimization risk by reducing both vulnerability and consequence, but at a cost

(a) Input values and initial risks

				Inputs and initial risk		
Node	T	V	C	Prevention cost	Response cost	Risk
Backup	0.5	0.5	1000	50	50	250
Grid	0.5	0.5	3000	600	600	750
Totals			4000	650	650	1000

(b) Results of risk minimization calculations

				After investment: $75 prevention; $50 response		
Node	T	V	C	Prevent allocation	Response allocation	Risk
Backup	0.5	0.32	702	5.77	3.85	111.7
Grid	0.5	0.32	2105	69.23	46.15	335.1
Totals		2807		75	50	446.8

(c) Results of risk maximization by an attacker with $125 thousand to invest. Defender adapts and changes allocation to minimize risk

				Attacker also invests $125; 3 unused			
Node	T	V	C	Prevent allocation	Response allocation	Attack allocation	Risk
Backup	0.82	0.18	389	13.3	10.24	21.8	57.4
Grid	0.56	0.33	2211	61.7	39.80	100.2	408.6
Totals			2600	75	50.04	122	466

to an *attacker–defender* competition with corresponding allocation of resources by each party.

Table 2.3c shows the results of a competitive game in which the attacker has $125 thousand in resources to increase threat where it will do the most harm. As shown in Table 2.3c, threat is increased to 82% probability for the backup asset and increased to 56% probability for the grid asset. Overall risk increases slightly to $466 thousand from $446.8 thousand.

MBRA's network optimization algorithm uses a Stackelberg[4] algorithm to obtain these results. First, the defender attempts to minimize risk by allocating resources to reduce V and C. Then, the attacker attempts to maximize risk by allocating resources to T. Generally, the attacker will increase T whenever the threat–asset–vulnerability triple gives the attacker a payoff in terms of risk. Conversely, the defender will attempt to decrease vulnerability everywhere, but the defender has a limited budget. Therefore, the defender is forced to leave low-consequence targets unprotected. Attacker and defender repeat their allocations until equilibrium is reached—neither attacker nor defender can improve on their optimizations. Equilibrium is not always possible, however, in which case MBRA stops after failing to reach a stalemate between defender and attacker. (In most cases the solution will oscillate between two or more equally minimum values.)

The game theory approach produces a value for T, given an attacker budget. Therefore, threat is an output value. But it is an output determined by assuming a rational actor always maximizes risk. What if the attacker has no interest in maximizing risk? Instead, it is entirely possible that a terrorist or criminal might act on opportunity or pure chance. In this case, T might better be obtained by other means. Such as described below under Bayesian belief networks.

The Transportation Security Administration (TSA) and the US Coast Guard have used more elaborate and sophisticated game theory to allocate resources. For example, the TSA GUARDS software attempts to minimize consequences due to a terrorist attack on airports by playing a two-party game with limited resources. GUARDS is described in more detail in Chapter 15. A similar approach has been used by the US Coast Guard to schedule cutter missions. ARMOR-PROTECT [3] minimizes risk by pitting a limited number of cutters against a much larger number of adversaries. It uses the RIN values obtained by MSRAM. PROTECT is described in more detail in Chapter 16.

[4]Stackelberg games honor German economist Heinrich von Stackelberg, author of *Market Structure and Equilibrium* (1934). https://en.wikipedia.org/wiki/Stackelberg_competition

2.8 BAYESIAN BELIEF NETWORKS

Rational actor models like Stackelberg assume an attacker thinks like an optimizer. But in reality, terrorists and criminals may be impulsive, irrational, or even stupid at times. This calls for an alternative method of prediction based on observable evidence. One such method is based on the 250-year old theory of Thomas Bayes (1701–1761)—a Presbyterian minister in England whose work was published only after his death in 1762. Even then his innovation was largely ignored until recently, because it is computationally intense—a chore best left to a computer.

Bayes believed in evidence instead of combinatorial analysis. He also conjectured that belief is conditional—it depends on mounting observations that either contribute to a belief or debunk it. He formalized his belief system in terms of *conditional probabilities*. For example, if the hospital modeled as a fault tree in Figure 2.3 is about to fail because its power is soon to be cut off, then it must be because the grid is damaged and the backup power is inadequate. Bayes formalized evidence as conditional probabilities strung together in a network as described in Appendix B.

Bayesian probability and its corresponding BN incorporate conditional probabilities to arrive at risk via observations rather than enumerated possibilities. A Bayesian probability asks, "What is the probability that the hospital H fails given the Grid is not operating and the Backup is not operating?" This language becomes more tortured as the number of conditions mount up. It also becomes more difficult to compute the conditional probabilities—called *beliefs*—as the size of the problem rises. For these reasons, a computerized BN is used to string together probabilities and conditional probabilities as shown in Figure 2.8.

Each node in a BN is a *proposition* that can be **true** or **false** or partially **true** or **false**. Partial truth is indicated by a probability—a number between zero and one. Probabilities are multiplied together as indicated above. But it is much easier to use a computer and BN software than arduously multiply fractions. More importantly, probabilities change from a number less than one to one, as evidence says the event has happened. In this sense, a BN is a real-time tool for estimating the likelihood of a future event as the events leading up to it are confirmed.

2.8.1 A Bayesian Network for Threat

Consider a simple threat analysis with the goal of estimating T for a bomb–bridge threat–asset pair. Intelligence analysts believe T is influenced by *capability* and *intent* as in the Coast Guard's MSRAM. Capability is measured by the degree of belief that a suspected terrorist can construct a bomb capable of blowing up a certain bridge. Intent is measured by the degree of belief that the suspect has performed surveillance on the bridge with the intention of blowing it

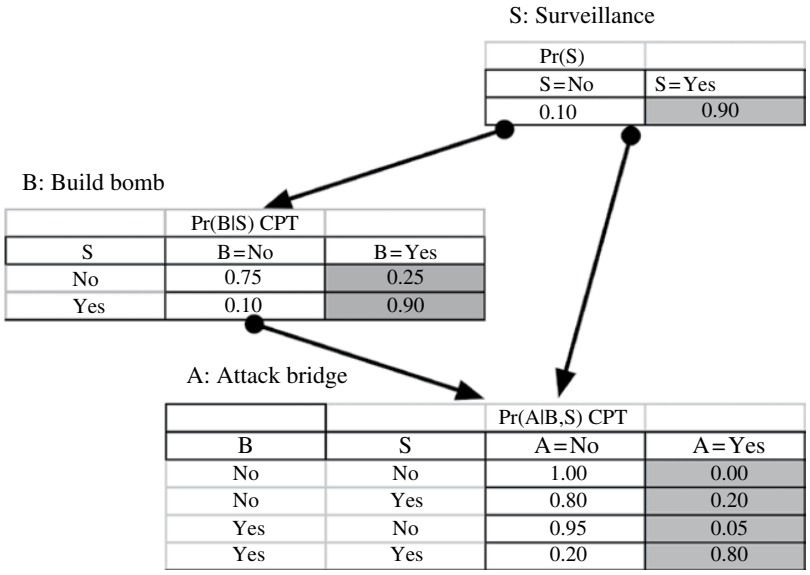


FIGURE 2.8 Bayesian network model of threat consists of three propositions: S, surveillance (intention); B, bomb-building (capability); and whether or not to A, attack a bridge. Beliefs are shown as probabilities in the conditional probability tables (CPT) of each proposition. Input beliefs are shaded.

up. Therefore, the evidence used to predict an attack on the bridge consists of surveillance, S, and bomb-building capability, B, as shown in Figure 2.8.

Every proposition has a truth table associated with it called the *conditional probability table* (CPT) that distills the degree of belief in every possible combination of preconditions. The CPT contains prior knowledge obtained from observation or experience and guesses based on supposition. These numbers quantify belief that a no or yes outcome will result from the preconditions. For example, setting S = Yes with probability 0.90 means we are 90% certain that surveillance indicates an intent to damage the bridge. Similarly, in Figure 2.8, probability of 0.25 means we are 25% sure that the suspect has the capability of building a bomb without previous surveillance, and 0.90 means we are 90% sure the capability exists when surveillance has been carried out. The attack CPT contains probabilities of attacking the bridge conditional on four preconditions. When no bomb capability or surveillance has been carried out, we believe the attack will not occur; when there is no bomb capability but surveillance has been carried out, the probability of an attack increases to 20%; and so forth.

The output from a BN changes when better, more accurate evidence is collected and entered in a CPT. For example, suppose a prior belief changes—we learn that the suspected terrorist has studied the bridge so surveillance changes from 90 to 100%. If we believe the terrorist has achieved bomb-building capability, the CPT for building a bomb increases to 100% also. These new priors are entered into the CPTs for surveillance and build bomb, and a new degree of belief in an imminent attack is calculated. In this case, belief in an eminent attack increases to 80%.

In general, a BN threat model like this can be reapplied every time new evidence arrives in the form of prior probabilities. The likelihood of the threat becoming real is an output value rather than an input. Of course, it can still be entered into a fault tree to allocate resources to minimize risk. But the main point of this analysis is to illustrate how threat can be calculated rather than entered as an input value only.

2.8.2 Predictive Analytics

Where do the estimates of belief come from in a BN such as the one shown in Figure 2.8? Typically, beliefs are mined from big data—statistics on past incidents and events. For example, in predictive policing, a metropolitan area might be divided into a checkerboard grid, and each square assigned numbers such as the number of traffic violations, number of shootings, number of terrorist attacks, and so on. These numbers are converted into estimates of the likelihood that an event such as a robbery or traffic accident will happen in each checkerboard square.

A more detailed BN might be based on contributing factors such as the number of domestic violence cases, the poverty level of a checkerboard square, unemployment levels, and so forth. Essentially, the BN is a model of the likelihood of one or more events occurring based on past histories. It is predictive only in the sense that the past is a predictor of the future.

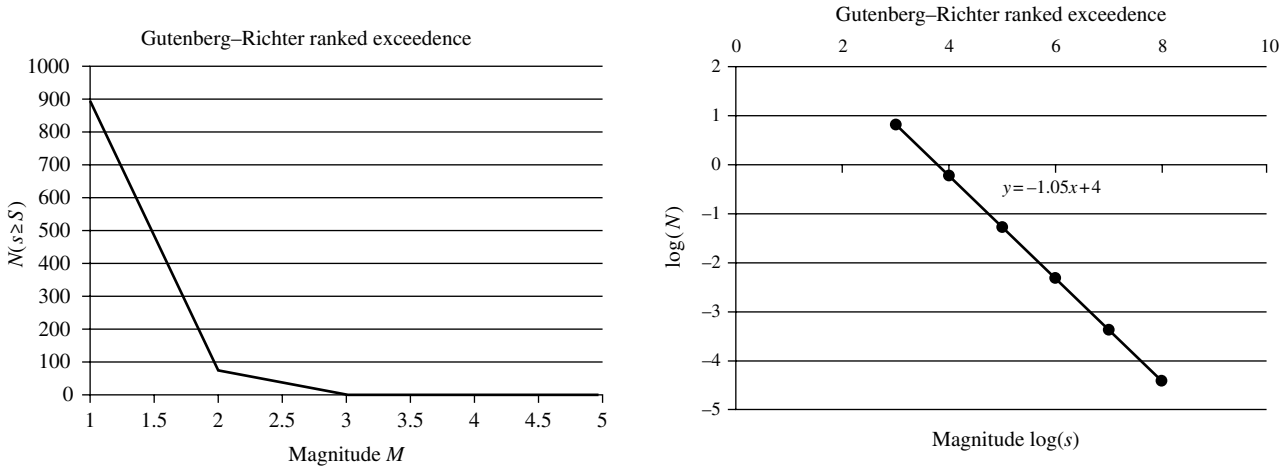


FIGURE 2.9 The Gutenberg–Richter law for relating the number of earthquakes of size S is an exceedence probability distribution that obeys a power law. The log–log plot is a straight line with slope equal to the fractal dimension of the power law.

2.9 RISK OF A NATURAL DISASTER

Mother Nature—not terrorists—causes most catastrophic events. In fact, terrorism is a low-risk hazard. Without a human perpetrator, the TVC model is useless for quantifying risk because there is no attacker and hence no threat component. But the expected utility formulation of risk is still useful, especially if the frequency of a hurricane, earthquake, flood, airplane accident, forest fire, or power plant meltdown is known. For example, given the probability of a forest fire, $\Pr(x)$ of size x, the expected loss is $\Pr(x)x$. Note that this is an entire table or graph of risks, because x can vary from zero to some maximum consequence. Therefore, risk is a *function* (see Appendix B for mathematical details).

2.9.1 Exceedence

Risk of a natural disaster is often quantified as an *exceedence* probability function rather than a single expected value, because insurers want to know the PML, so they can adjust premiums accordingly. PML risk is based on exceedence probability, which is the likelihood of an event of size C or larger. The famous Gutenberg–Richter scale for measuring the size of an earthquake is an example of an exceedence probability distribution, EP (see Fig. 2.9).

Note that a typical EP is *long-tailed*, meaning that the curve drops rapidly from an initial high value to a low value, left to right. The longer the tail, the heavier it is; hence long-tailed distributions as illustrated in Figure 2.9 are also called heavy-tailed distributions. Such a distribution means that it is much more likely that an event near zero on the x-axis will occur than an event far to the right side of the graph. Furthermore, a long-tailed distribution declines more slowly along the $x \gg 0$ axis, indicating that likelihood still exists even at the extreme right side of the graph.

A classical long-tailed distribution is mathematically equivalent to a power law. While other mathematical functions qualify as long-tailed, power laws are simple to represent as a mathematical function and actually occur in nature. As it turns out, the EP curve of most hazardous events is a *power law* with a long tail when plotted on an x–y-axis, but a straight line when plotted on a log–log-axis chart.[5] Figure 2.9 illustrates both—a long tail when plotted on a linear graph and a straight line when plotted on a log–log graph.

Power laws are fully described by a single number called a fractal dimension, which is simply the slope of the EP plotted on log–log scales as shown in Figure 2.9. A low value of fractal dimension corresponds to a relatively horizontal flat EP curve. A high value corresponds to a relatively rapid vertical drop in the EP, while a low value corresponds with a relatively low rate of decline or even a nearly flat curve. The fractal dimension of Figure 2.9 is 1.05, because the slope of the log–log plot is (−1.05).

2.9.2 EP vs. PML Risk

PML risk is calculated from an exceedence probability curve EP by multiplying values along the x-axis by corresponding values along the y-axis as illustrated in Figure 2.10. That is, PML risk is the product of the likelihood of an event of consequence x or larger and EP. The dotted lines of Figure 2.10 illustrate the difference between EP and PML.

Threats and hazards that increase PML risk without bound, as in Figure 2.10a, are considered high risk, while Figure 2.10b shows a low-risk threat or hazard. Note that a low-risk threat or hazard reaches a peak, which is the maximum PML risk, often simply called PML risk. Maximum

[5]For mathematical details, see Appendix B.

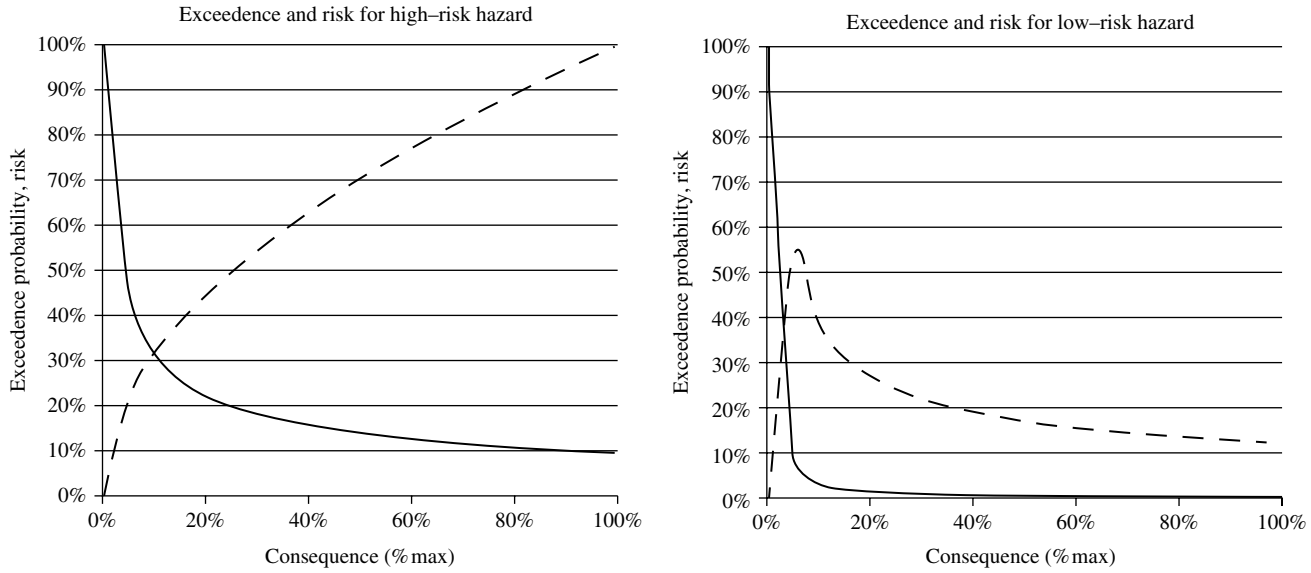


FIGURE 2.10 Long-tailed exceedence probability curves: one for high-risk hazards and the other for low-risk hazards.

PML risk represents the greatest exposure for an insurance company and is often called maximum probable loss in contrast to PML.

Most threats and hazards described in this book are low risk in the sense of PML and reach a peak as shown in Figure 2.10b. The magnitude of the peak of the dotted line in Figure 2.10b is a compact measure of risk we will use throughout this book. It represents the likely worst-case scenario for a threat or hazard.

Exceedence probability EP is a measure of the likelihood of a worst-case event, while PML is a measure of the expected loss due to a worst-case event.

2.10 EARTHQUAKES

Perhaps the most famous *ranked exceedence* curve is the Gutenberg–Richter scale for measuring the size of earthquakes. According to the USGS, "The Richter magnitude scale was developed in 1935 by Charles F. Richter of the California Institute of Technology as a mathematical device to compare the size of earthquakes. The magnitude of an earthquake is determined from the logarithm of the amplitude of waves recorded by seismographs. Adjustments are included for the variation in the distance between the various seismographs and the epicenter of the earthquakes. On the Richter scale, magnitude is expressed in whole numbers and decimal fractions. For example, a magnitude 5.3 might be computed for a moderate earthquake, and a strong earthquake might be rated as magnitude 6.3. Because of the logarithmic basis of the scale, each whole number increase in magnitude represents a tenfold increase in measured amplitude; as an estimate of energy, each whole number step in the magnitude scale corresponds to the release of about 31 times more energy than the amount associated with the preceding whole number value."[6]

The magnitude of an earthquake, M, is an expression of the amount of energy released by a tectonic shift. Richter observed that the number of earthquakes in excess of a certain size obeys a power law with a fractal dimension close to 1.0. The *number* of earthquakes falls on a straight line when plotted on a log–log graph. Generally, exceedences are plotted on log–log graphs so we can obtain the fractal dimension from the slope of the straight line. (It is also used to determine if exceedence is truly a power law.)

2.11 BLACK SWANS AND RISK

The term *black swan* achieved mainstream status with the publication of Taleb's highly successful book, *The Black Swan* [4]. Taleb uses the colorful phrase to describe highly unlikely and yet highly consequential events that surprise and astonish us. He addresses a major limitation of EUT by asking, "what is the definition of risk when Pr(failure) approaches zero and consequence approaches infinity?" After all, the product of zero times infinity is mathematically undefined. What does it mean? (The terrorist attack of 9/11 is an example: its likelihood approached zero and its consequence approached infinity. It seems to render EUT useless.)

This paradoxical question is easily answered when true exceedence probability is a power law and PML risk is taken to the limit. A black swan event occurs with proba-

[6]http://earthquake.usgs.gov/learn/topics/richter.php

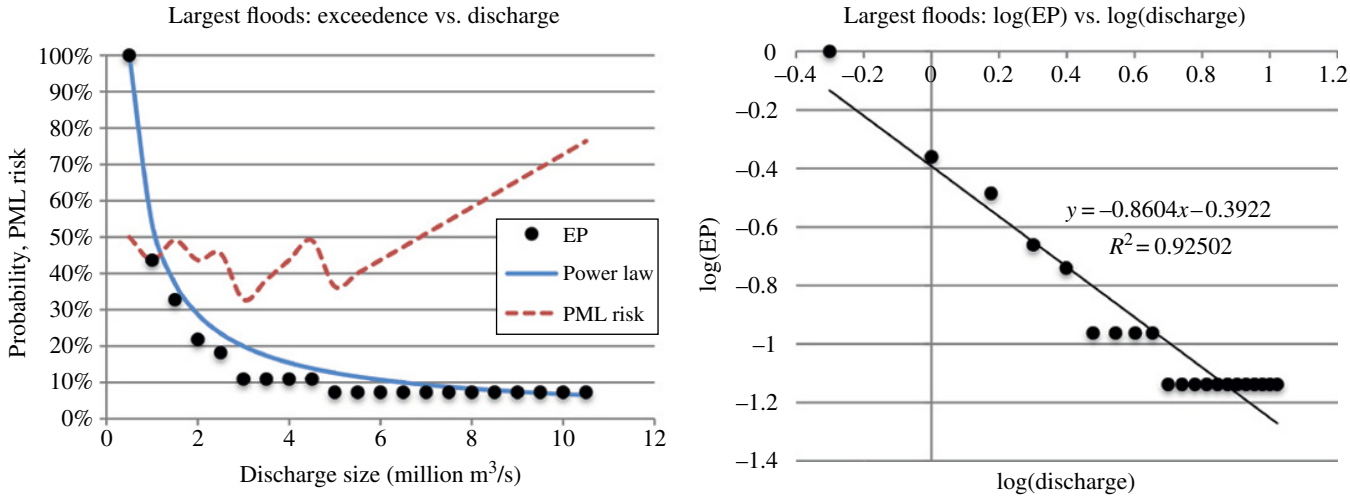


FIGURE 2.11 Exceedence probability—in x–y coordinates and log(x)–log(y) coordinates—of the largest known floods in history reveals them to be high risk.

bility equal to the right-hand-most point in the true exceedence probability function; see the rightmost points in Figure 2.10. This is where the extremely large-consequence, extremely small likelihood incidents lie on the graph. Taleb's conundrum is resolved by observing what happens to PML risk as x-axis consequence increases without bound.

When consequence increases without bound, PML also increases without bound, if the fractal dimension of the hazard's true exceedence probability curve is less than 1. Otherwise, PML risk drops to zero as consequence increases without bound. We call the unbounded PML risk curve *high risk* and the bounded PML risk curve *low risk*. Whether a threat or hazard is considered high or low risk depends entirely on the value of fractal dimension:

High risk: fractal dimension < 1.0
Low risk: fractal dimension > 1.0

The shape of these curves is completely determined by the fractal dimension of the exceedence probability curve. High-risk hazards have longer-tailed exceedence probability curves, and low-risk hazards have shorter-tailed curves. The fractal dimension of the power law EP curve tells us the most important fact about a hazard—its qualitative degree of risk. It also solves Taleb's conundrum.

2.12 BLACK SWAN FLOODS

Jim O'Connor is a scientist with the U.S. Geological Survey Water Science Center in Portland, Oregon. He is a native of the Pacific Northwest "long interested in the processes and events that shape the remarkable and diverse landscapes of the region," according to his home page.[7] Along with colleague John Costa, the two studied the largest known floods over the past 1.8 million years (the Quaternary Period) and found them to be high-risk hazards [5]. The exceedence probability graphs in Figure 2.11 were obtained from data composed by O'Connor and Costa.

According to their research report, "The largest known floods of the Quaternary Period had peak discharges of nearly 20 million cubic meters/second and resulted from breaches of glacial-age ice dams that blocked large midcontinent drainage systems during ice ages. Most of the other largest documented floods resulted from breaches of other types of natural dams, including landslide dams, ice dams from smaller glaciers, releases from caldera lakes, and ice-jam floods. Only 4 of the 27 largest documented floods were primarily the result of meteorological conditions and atmospheric water sources. However, if only historic events are considered, the proportion of large meteorological floods still rises to only 4 of 10."[8]

The fractal dimension of these black swan floods is less than 1.0 as shown in the log–log graph of Figure 2.11. Therefore, black swan floods are high risk. The dashed PML risk line is unbounded as size increases. In other words, PML risk increases without bound as consequence, measured in volume of water discharged, increases. Recall that the dotted line in Figure 2.11 is obtained by multiplying x-axis values by y-axis values to get PML risk.

[7] https://profile.usgs.gov/oconnor/
[8] http://pubs.usgs.gov/circ/2004/circ1254/pdf/circ1254.pdf

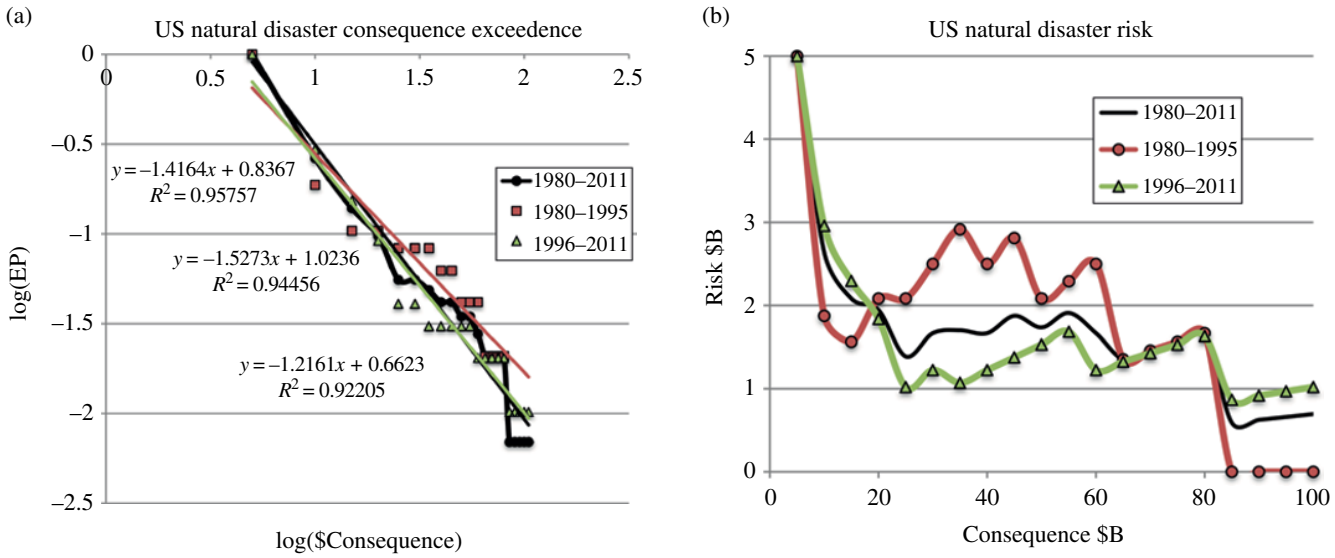


FIGURE 2.12 Exceedence probability of financial consequences from US natural disasters 1995–2011 is shorter-tailed than during the 15-year period 1980–1995. (a) Fractal dimensions of consequence exceedence for 1980–1995 is 1.22; 1996–2011 is 1.42; and 1980–2011 is 1.53. (b) Risk profiles of natural disasters 1980–2011, 1980–1995, and 1996–2011 suggests risk is lower now than the 1980–1995 period, except for the last decade (2000–2011).

2.13 ARE NATURAL DISASTERS GETTING WORSE?

US natural disasters costing in excess of one billion dollars—hurricanes, forest fires, droughts, floods, and so on—during the 15-year period 1980–1995 occurred with a frequency and size that yielded a fractal dimension of $q = 1.22$ (see Fig. 2.12a). They were low-risk hazards because $q > 1$, but during the 15-year period 1995–2011, natural disasters *decreased* in frequency and size so that the fractal dimension of consequence exceedence rose to 1.42. In other words, the tail of consequence exceedence is getting shorter. This suggests less risk, but note the cross over in risk that occurred around 2000.

The risk profiles of Figure 2.12b show an unexpected upturn in risk after the year 2000. The risk profile of the 1995–2011 period crosses over the 1980–1995 tail, suggesting a reversal of fortunes. For most of the 30-year period, natural disaster hazards were becoming less lethal. From Figure 2.12b it appears that natural disasters are suddenly getting worse after decades of being lower risk than the 30-year trend. Why is this reversal of fortunes happening?

Global warming is one common explanation of the sudden reversal in natural disaster risk. As the temperature of the earth rises, it alters Earth's ecosystem in nonlinear ways not fully understood by science. However, we have 160 years of temperature readings to study and determine if there is a connection between the slowly rising mean temperatures and weather-related hazards such as superstorms, hurricanes, droughts, and floods. Can we correlate global warming with the risk of natural disasters increasing since 2000?

2.14 BLACK SWAN AL QAEDA ATTACKS

Paul Pierre Levy (1885–1971) was a French mathematician and educator that studied a type of random walk found in nature—and subsequently named after him—the *Levy flight* or *walk*. Levy flights and walks are distances traveled by animals and people containing waypoints separated by a distance that obeys a power law distribution. For example, foraging animals in the forest, shoppers in the mall, vacationers in Disney World, and contagions in a global pandemic hop, skip, and jump according to a Levy flight pattern.

The author studied Levy flights of the al Qaeda terrorist organization, the SARS pandemic, and shoppers in Manhattan, New York, and found they all obey Levy flights and walks. Furthermore, in many cases, terrorist attacks and other catastrophic events also obey Levy flights in time as well as distances between subsequent attacks. Both elapsed time and distances between subsequent events obey power laws—suggesting that al Qaeda attacks and epidemics are similar in complex behavior.

Consider the dramatic example of exceedence probability and fractal dimensions obtained by plotting consequences (deaths), *displacement* (miles), and *elapsed times* (days) between subsequent terrorist attacks in Figure 2.13 [6, 7]. Displacement is defined as the distance between subsequent attacks, and elapsed time is defined as the time

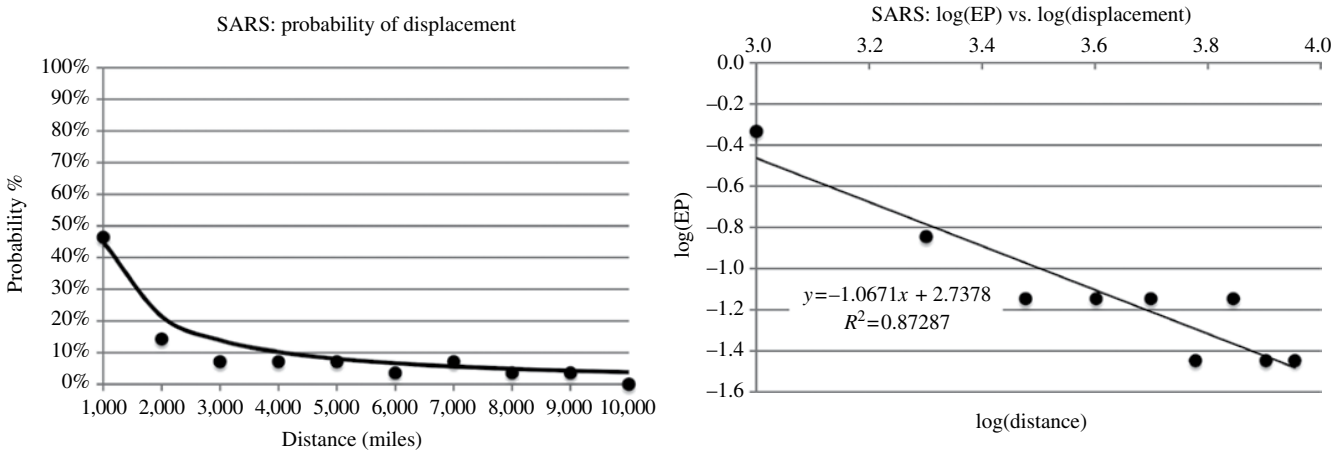


FIGURE 2.13 Fractal dimension of the Levy flight of SARS as it spread to 29 countries was equal to 1.07.

interval between subsequent attacks. Exceedence is the true exceedence probability obtained by summing the frequency of events from the black swan end of the graph to its *x*-axis value. In each case, the exceedence probability distribution equals the probability of an event of consequence, distance, or elapsed time greater than or equal to number of deaths, miles, or days.

They all form Levy flights of different fractal dimensions. In terms of deaths, al Qaeda attacks are low risk, because the exceedence probability curve fits a power law with fractal dimension greater than one (1.33). This means the PML risk initially rises but then falls to zero as consequence rises. The number of deaths caused by al Qaeda on 9/11 is a black swan with exceedence odds of 1 in 999, corresponding with PML risk of 2.7 deaths.

Levy flight analysis of distance suggests that most al Qaeda attacks were relatively local, because of the lopsided power law fit to the exceedence probability curve of Figure 2.13b. But the fractal dimension suggests the opposite—al Qaeda has a relatively long reach, because of the long tail of the curve. A fractal dimension less than one confirms this. The terrorist group operates on a global scale—the 9/11 attacks in New York, Virginia, and Pennsylvania were more than 5000 miles from the previous attack.

Levy flight analysis of the elapsed time between subsequent attacks during this interval (1993–2009) suggests short time intervals because the fractal dimension of the time exceedence curve is greater than one. In fact, if bin Laden had not been killed and al Qaeda largely defeated, the probability of the next attack occurring before *x* days in the future would be 1 – EP(*x*), from Figure 2.13c. For example, the probability of a subsequent al Qaeda attack within the next 360 days is 0.94, according to the power law with fractal dimension of 1.4. (At the time this was written, several attacks occurred that were not included in the initial analysis.)

2.15 BLACK SWAN PANDEMIC

Figure 2.14 shows the probability distribution for the SARS mini-pandemic that originated in China in 2002 and was largely extinguished 9 months later. It could have been different. SARS flared up in November 2002; infected 8422 people, killing 916 in 29 countries; and then faded out by June 2003. SARS potentially could have spread to every corner of the globe, except it did not. Why not?

Probability distributions, Levy flights, and PML risks apply to biological threat–asset pairs such as global pandemics—or threats of pandemic—just as they apply to other natural disasters. For most of the 80 years following the 1927 Kermack–McKendrick mathematical model of epidemics, the world assumed germs bounced from host to host like a random golf ball [8]. The chance of contracting the black plague was an equal-opportunity disease—everyone had an equal chance of contracting it. But modern *network science* replaced the Kermack–McKendrick model with a modern one: contagious germs take Levy flights from host to host. They do not bounce around completely at random. Instead, their probability of spreading a certain distance obeys a power law. Pandemic diseases travel along Levy flights defined by social networks, and social networks are not random.

Global pandemics spread through networks created by people socially connecting to one another. This social network forms pathways for Levy flights—literally airline flights from one country to another. While SARS began in a backwoods region of China, it was soon carried to a hotel in Kowloon where infected Doctor Liu Jianlun waited for the elevator on the ninth floor. The doctor had recently attended a patient with a strange respiratory disease. Next to him were three Canadians, a man and woman from Hong Kong, and an American businessman. The Canadians, Chinese, and American got on airplanes and flew to other countries. Doctor

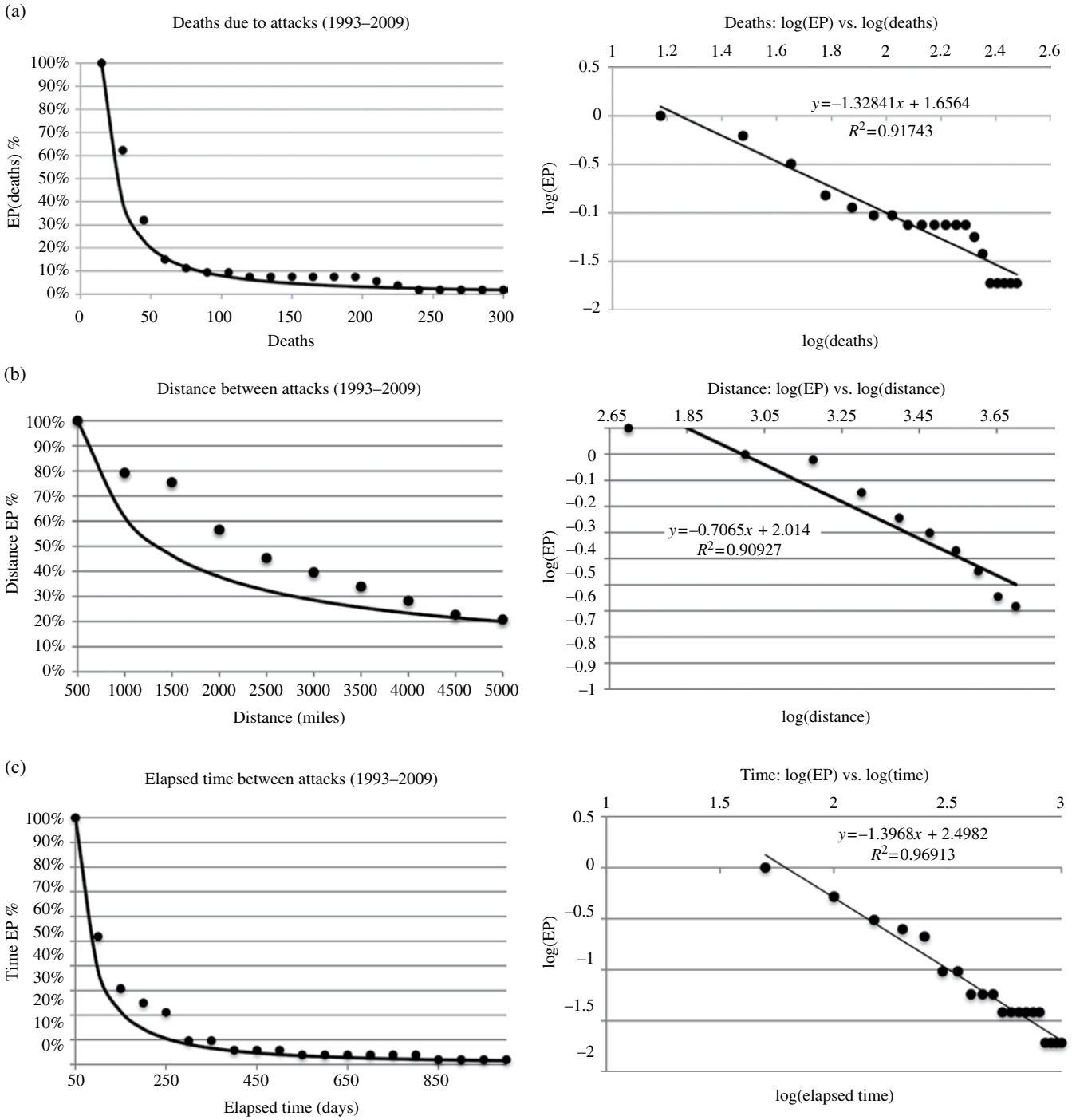


FIGURE 2.14 The exceedence probability distributions for consequence, distance, and elapsed time between subsequent al Qaeda attacks during the period 1993–2009 are Levy flights. Source of data: http://www.infoplease.com/ipa/A0884893.html. (a) Fractal dimension of exceedence probability versus number of deaths due to al Qaeda attacks equals 1.33 (low risk). (b) Fractal dimension of Levy flight of al Qaeda attacks versus distance (miles) between subsequent attacks equals 0.71 (long distances). (c) Fractal dimension of Levy flight of al Qaeda attacks versus elapsed time (days) between subsequent attacks equals 1.40 (short time intervals).

Liu soon died from the strange respiratory disease, and the others spread the disease to other countries. Incredibly, the lift on the ninth floor was the epicenter of a social network that spread the strange disease to 29 countries.

The question is *not* why did SARS travel the globe—the question is, why did it stop? The answer is social networking SARS foraged its way over one-third of the globe and then stopped in its tracks by a combination of alert public health

experts and the fractal dimension of its probability distribution. According to researchers Hu, Luo, Xu, Han, and Di, pandemics die out if the *fractal dimension* of the associated *probability distribution* is less than two. The probability distribution is formed by the frequency of contagion, which in turn is determined by the shape of the social network. And the social network is shaped by air travel and face-to-face contact. Hu *et al*. claim, "the epidemic is liable to disappear if there are more long-distance connections than short ones" [9]. In other words, in addition to a fast-acting public health sector, the social network stretched too thin across the globe, which weakened and eventually killed SARS. In an age of global air travel, long flights are safer than short flights.

The fractal dimension of the SARS probability distribution power law is far less than the required tipping point of two. Therefore, SARS burned out partially because the human carriers traveled too far. In addition, health professionals were able to stamp out the disease in each distant country before it started spreading again. The combination of rapid response and global air travel defeated SARS.

There are two fundamental kinds of epidemics. The first kind, susceptible–infected–recovered (SIR), describes a population of people that are initially susceptible to a disease, then infected with a certain probability, and finally either recover or die, so they are no longer susceptible. SIR diseases eventually die out because they either kill their hosts or run out of victims.

The second kind of disease is known as susceptible–infected–susceptible (SIS). It spreads throughout a population of initially susceptible individuals that either die or recover and become susceptible again. An SIS population can sustain a contagion forever, if conditions are right for recurrence of the disease. SARS was suspected of being SIS, which means it can flare up again. If the initial victims at ground zero stay in their own country, the contagion could grow to an enormous size before public health officials get ahead of the disease. But if the initial victims travel long distances more than short distances, the disease is likely to be controlled if the fractal dimension is less than two.

Chapter 14 examines the impact of the commercial air travel network on the spread of contagions and recommends a number of countermeasures for reducing the spread of highly contagious diseases. The solution involves a combination of quick response and complexity theory remedies (blocking nodes). As it turns out, the air travel network can be exploited to stop pandemics.

2.16 RISK AND RESILIENCE

Risk is not a single number. It is an entire function as illustrated by Figure 2.10. Resilience, however, is quantifiable as a single number—it is proportional to the fractal dimension of a critical infrastructure exceedence probability function.[9] Resilient threat–asset pairs have higher fractal dimensions, which means shorter-tailed exceedence probability distributions. So resilience of a threat–asset pair network is proportional to the fractal dimension of the hazards applied to the threat–asset pair. Resilience increases as fractal dimension increases. The way to reduce risks due to hazards like hurricanes and terrorism is to shift the exceedence probability from high to low risk—by shortening its tail. Risk reduction and antifragility strategies come from bending the exceedence curve downward (increasing its fractal dimension) to make the tail shorter and thinner.

> ***Risk/Resilience Strategy***: *Threat–asset pair risk is reduced and resiliency improved by increasing the fractal dimension of the threat–asset pair's exceedence probability function. Resilience is proportional to fractal dimension—higher is better.*

Table 2.4 lists a number of common hazards and their fractal dimensions. They are divided into low- and high-risk categories according to their fractal dimensions. It may be surprising to some that terrorism is a low-risk hazard. It may also come as a surprise that fractal dimension varies so widely from hazard to hazard.

Consequence is measured in different ways in Table 2.4. For example, the consequence of earthquakes is measured in area, deaths, and financial loss. When measured in area, earthquakes are considered low risk. But when measured in deaths and financial loss, they are considered high risk. Why?

There are more threats and hazards than we have funds to prevent. Therefore, optimal resource allocation should be based on ROI—a cold analytic method of deciding what is protected and what is not. In addition, a diversified portfolio of critical infrastructure investments is the best strategy, because of diminishing returns. Each asset or system in the portfolio should be resourced optimally, but not beyond some predetermined ROI. To spend limited funds beyond a certain ROI is wasteful. There is not enough money to protect every CIKR asset or system, so optimal allocation is essential.

Risk and resilience assessment as described here is a partial solution to the problem of measuring the ability of a CIKR to resist, absorb, and adapt to adversity, as the definition of resilience demands. Risk is expected loss and therefore quantifies the impact of adversity on a CIKR. Fractal dimension measures one kind of ability to absorb adversity—the ability to absorb system failures. It does not quantify the robustness—the ability to function under stress. For example, a robust system may employ redundancy to adapt

[9]Later we will learn that resilience is proportional to the product of probability of cascade failure and spectral radius.

TABLE 2.4 Hazards may be classified as high- or low-risk hazards by the fractal dimension of the power law that best fits the hazard's exceedence probability curve. Fractal dimensions less than 1.0 indicate high-risk hazards, while fractal dimensions greater than 1.0 indicate low-risk hazards

Hazard	Consequence	Fractal dimension
		Low risk
S&P500 (1974–1999)	$Volatility	3.1–2.7
Large fires in cities	$Loss	2.1
Airline accidents	Deaths	1.6
Tornadoes	Deaths	1.4
Terrorism	Deaths	1.4
Floods	Deaths	1.35
Forest fires in China	Land area	1.25
East/West power grid	Megawatts	1
Earthquakes	Energy, area	1
Asteroids	Energy	1
Pacific hurricanes	Energy	1
		High risk
Hurricanes	$Loss	0.98
Public switched telephone	Customer-minutes	0.91
Largest known floods	Discharge	0.81
Forest fires	Land area	0.66
Hurricanes	Deaths	0.58
Earthquakes	$Loss	0.41
Earthquakes	Deaths	0.41
Wars	Deaths	0.41
Whooping cough	Deaths	0.26
Measles	Deaths	0.26
Small fires in cities	$Loss	0.07

to adversity. When one component fails, a robust system switches to a backup component and continues to function.

In the following chapters, robustness and redundancy will be shown to *increase* risk and *decrease* resiliency, in some cases. That is, a robust and redundant system may make risk and resilience worse, rather than better. This may seem counterintuitive, but remember that as a system becomes more complex, it also becomes more likely to trigger a cascade of faults ending in collapse. Risk assessment is only one tool used to evaluate CIKR performance. Other measures such as redundancy, surge capacity, recovery time, and cost play a role, too.

2.17 EXERCISES

1. Expected utility theory was invented by:
   a. Blaise Pascal
   b. Daniel Bernoulli
   c. Norman Rasmussen
   d. Ralph Nader
   e. Laplace
2. A threat–asset pair is:
   a. Fundamental building block of CIP risk
   b. Natural disasters and buildings
   c. Floods and buildings
   d. The definition of hazard
   e. PRA
3. Which one of the following is true?
   a. Threat is a bomb, cyber exploit, hurricane, or accident.
   b. Vulnerability is a flaw.
   c. Consequence is cost.
   d. Risk is expected loss.
   e. Exceedence is a measure of risk.
4. Fault tree vulnerability is:
   a. The probability of tree failure
   b. The probability of a threat–asset pair failure
   c. The probability of an AND event
   d. The probability of an OR event
   e. All of the above
5. Return on CIKR investment is:
   a. How much it cost to reduce risk to zero
   b. The ratio of risk reduction to investment
   c. How much vulnerability is reduced
   d. How much consequence is reduced
   e. The cost of response
6. One major deficiency in PRA is:
   a. It is too expensive.
   b. It suffers from diminishing returns.
   c. It cannot be used to minimize risk.
   d. It assumes threat is always an input variable.
   e. It only works for nuclear power plants.
7. Stackelberg competition (in MBRA) is used to:
   a. Estimate threat T
   b. Estimate reduced vulnerability V
   c. Estimate reduced consequence C
   d. Optimize risk
   e. All of the above
8. Bayesian network analysis is based on Bayes' theory of:
   a. Conditional probabilities as beliefs
   b. The chain rule
   c. The calculation tree
   d. Optimization of resources
   e. Allocation of resources
9. Exceedence probability is:
   a. The probability of an earthquake
   b. The probability an event will equal or exceed a certain level
   c. The math behind Gutenberg–Richter scale
   d. The math behind the largest floods
   e. The rank of the top 100 movies of all time
10. A power law is a probability distribution with:
    a. A long tail
    b. A black swan
    c. A ranked exceedence probability

d. A high-risk incident
e. An exponential decline versus consequence

11. In this chapter, resilience is measured by:
    a. Ranked exceedence
    b. True exceedence
    c. Fractal dimension
    d. PML risk
    e. The resilience triangle
12. High-risk hazards have:
    a. High consequences
    b. High vulnerability
    c. High cost of recovery
    d. Fractal dimension, $q < 1$
    e. Long recovery times
13. According to some scientists, SARS died out quickly because:
    a. Its Levy flight took too many long steps.
    b. Public health officials in China responded quickly.
    c. It started with only a handful of people.
    d. Viruses cannot jump from animals to humans.
    e. The airline companies filter airplane air.
14. What are the units of measurement of risk?
    a. The same units as consequence
    b. Dollars
    c. Casualties
    d. Time
    e. Cubic meters per second
15. Which of the following is *not* a probability?
    a. Threat
    b. Vulnerability
    c. Exceedence
    d. Binomial distribution
    e. Consequence

2.18 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. Most risk-informed decision-making frameworks depend on qualitative and/or quantitative analysis of data such as T, V, and C. But these are not the only values that might be used to calculate risk. Propose and justify alternative measures for calculating risk.
B. Fault tree analysis is derived from the nuclear power industry. Alternatives to this model were described in this chapter. What is decision tree analysis and how does it differ from the methods described here?
C. Why do you suppose so many catastrophes obey a power law in terms of exceedence? Include the notion of scale-free properties of fractals in your answer.
D. The section on high-risk versus low-risk catastrophes suggests different strategies for high-risk versus low-risk events. In particular the trade-off between prevention and response depends on the extreme nature of high-risk events. What other strategies might be employed? Suggest an alternate strategy for making investments in prevention versus recovery, and explain why your strategy is superior to the high/low-risk strategy proposed here.

REFERENCES

[1] Hutchins, E. M., Cloppert, M. J., and Amin, R. M., *Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains*. Bethesda, MD: Lockheed Martin Corporation. https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf

[2] LCDR Brady Downs. The Maritime Security Risk Analysis Model, *MSRAM Proceedings*, 64, 1, Spring 2007, pp. 36–38.

[3] Shieh, E., An, B., Yang, R., Tambe, M., Baldwin, C., DiRenzo, J., Maule, B., and Meyer, G. Protect: A Deployed Game Theoretic System to Protect the Ports of the United States. *International Conference on Autonomous Agents and Multiagent Systems* (AAMAS), Valencia, Spain, 2012. Available at http://teamcore.usc.edu/tambe. Accessed June 25, 2014.

[4] Taleb, N. N. *The Black Swan: The Impact of the Highly Improbable*. New York: Random House, 2007.

[5] O'Connor, J. E. and Costa, J. E. *The World's Largest Floods, Past and Present-Their Causes and Magnitudes*. U.S. Geological Survey Circular 1254. Reston, VA: U.S. Geological Survey, 2004, pp. 13. Available at http://pubs.usgs.gov/circ/2004/circ1254/. Accessed June 25, 2014.

[6] Clauset, A., Young, M., and Gleditsch, K. S. On the Frequency of Severe Terrorist Events, *Journal of Conflict Resolution*, 51, 1, 2007, pp. 58–87.

[7] Lewis, T. G. *Bak's Sand Pile*, 2nd ed, Monterey: AgilePress, 2011, pp. 153.

[8] Kermack, W. O. and McKendrick, A. G. A Contribution to the Mathematical Theory of Epidemics, *Proceedings of the Royal Society of London*, 115, 1927, pp. 700–721.

[9] Hu, Y., Luo, D., Xu, X., Han, Z., and Di, Z. Effects of Levy Flights Mobility Pattern on Epidemic Spreading under Limited Energy Constraint, arXiv:1002.1332v1 [physics.soc_ph] pp. 1–5. Available at http://arxiv.org/abs/1002.1332. Accessed, February 5, 2010.

3

THEORIES OF CATASTROPHE

The study of critical infrastructure protection, and homeland security in general, is more than the study of terrorist attacks, natural disasters, and accidents. In fact, there is a rich theoretical basis for the study of catastrophes, complex systems, and the relationship between complexity and failure of CIKR systems. This chapter traces the historical development of the three major theories: Perrow's *normal accident theory* (NAT), Bak's theory of *punctuated equilibrium*, and the more recent insights obtained by applying complex adaptive systems theory to CIP.

Biological and ecological systems are among the most complex systems in existence. It is not surprising, then, that principles observed in biological systems also apply to human-made CIKR systems. These principles are often expressed in terms of paradoxes and parables such as the tragedy of the commons (TOC), paradox of enrichment (POE), paradox of redundancy (POR), and competitive exclusion principle (CEP). The combination of NAT, complexity theory, and a handful of biological principles form the basis of a comprehensive and modern theory of catastrophes.

This chapter surveys the three theories and develops new measures and insights into complex CIKR systems by borrowing principles from biology as follows:

- *Normal accident theory*: Charles Perrow's 1979 theory says extreme events occur when two or more failures occasionally come together in an unexpected way, are accelerated and increased in severity if the system is *tightly coupled*, and grow to catastrophic proportions when the system has *catastrophic potential*. But Perrow does not elaborate on the definition of *catastrophic potential*.
- *Bak's punctuated equilibrium theory*: Per Bak, Chao Tang, and Kurt Wiesenfeld observed catastrophic collapses of a hypothetical sand pile in the mid-1980s in an experiment that became known as the *BTW experiment*, aka *sand pile* experiment. The sand pile experiment became a metaphor for simple systems that behave in complex ways. The BTW experiment formed the basis of modern complexity theory and led Bak to formulate a more general theory he called *punctuated equilibrium*. Incidents impacting complex systems such as the electric power grid, air transportation, and Internet are bursty as they occur according to long-tailed probability distributions. Bak attributed this bursty or punctuated behavior to the buildup of *self-organized criticality* (SOC) that inevitably builds up in complex systems. SOC is Perrow's catastrophic potential and explains why consequences magnify as a fault spreads through a CIKR system.
- *Self-organization*: Bak's SOC generally stems from increasing efficiency and optimizing system performance, which eliminates redundancy and surge capacity from CIKR systems as their architectures evolve from fragmented and somewhat random structure to integrated, linked, and highly structured systems. Self-organized systems are typically nonredundant, nonsurge capable, single-point-of-failure systems with bottlenecks, overly concentrated assets, and inadequate

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

backup capacity. SOC equates with fragility or antiresiliency and is the main source of fragility. Consequences are larger when SOC is greater, in general.

- *TOC*: In addition to SOC, complex systems can contain the seeds of their own destruction due to nonlinearities that become apparent only when the system is under stress. For example, *sustainability*—the ability of a complex system to continue indefinitely under a variety of conditions—is one of the most common victims of nonlinearity in the behavior of a complex CIKR system. Sustainability—or the lack of it—is captured in a simple parable of a shared resource called a *commons*. The TOC parable says that a system may not be sustainable if the actions of a self-interested *predator* overwhelm the capacity of an underlying commons—called the *prey*—and depletes the commons. When the predator exceeds the *carrying capacity* of the commons, the system becomes unstable and is likely to collapse. Collapse comes unexpectedly because of an inherent nonlinear relationship between the load placed on the system by predator and the inherent carrying capacity of the commons. TOC is used as a test, here, to determine if a complex CIKR system is sustainable.
- *POE*: A corollary to TOC is another nonlinear side effect of complexity called the paradox of enrichment—a behavior that occurs when a complex system becomes unstable due to an enrichment that exceeds the organic carrying capacity of the commons. Enrichment of the infrastructure destabilizes the predator–prey balance by overshooting the ecosystem's carrying capacity. Enrichment may appear to enhance the system for a short time, but ultimately, it causes collapse. POE is used as a test of s complex CIKR system's ability to absorb "too much, too fast." It is a behavior that is especially prevalent in financial systems.
- *Minsky moment*: Over exuberant investors may experience a Minsky moment when the economy is overheated by enrichment of certain sectors of the economy as happened in the 2000s, resulting in the 2008 financial meltdown in the United States. A Minsky moment is a type of POE repeated by the banking and finance sector. Braess's paradox is a second example of POE that says; adding lanes to a congested highway (enrichment) slows traffic even more (paradox). In some CIKR systems, such as highly centralized information technology departments, POE threatens to collapse the IT function because enrichment exceeds the IT department's carrying capacity.
- *State space*: The stability of a commons or CIKR system can be studied by plotting predator against itself or against its prey and observing the trajectory of the graph. This plot is called a state space diagram, and the pattern it traces out reveals potential instability of a complex CIKR. A system is stable if its state space trajectory is bounded and nonzero, metastable if the trajectory circles indefinitely, or unstable and chaotic if the trajectory is erratic or unbounded. A *fixed point* in state space is point in a trajectory where the system enters and never leaves. The state space diagram is used to analyze a CIKR to determine if it is stable, metastable, or unstable and likely to fail.
- *CE)*: The CEP says that competitive ecosystems tend to eliminate all but one competitor, because sooner or later, one competitor gains a small advantage over all others and grows faster and becomes fitter than all others. This leads to a monopoly, in general, which reduces redundancy and diversity. CEP diminishes resilience, largely because monopolies are optimized organizations that tend to build optimized (profitable) systems. In general, critical infrastructure systems abhor competition and tend to become monopolies, which is a form of "putting all your eggs in one basket."
- *Preferential attachment*: Preferential attachment is the most common form of self-organization that leads to SOC and CEP. In practice, preferential attachment creates concentrations of assets, bottlenecks, and single points of failure in CIKR. The Internet is currently undergoing restructuring because of preferential attachment—a hub-and-spoke architecture is emerging due to economics and regulation. Power grids and transportation systems have reached high levels of SOC due to decades of preferential attachment. Wherever preferential attachment is at work, the resulting system is likely to be vulnerable, because of a critical hub, essential bottleneck, or "weakest link."

3.1 NORMAL ACCIDENT THEORY (NAT)

Charles Perrow is perhaps the first modern person to study catastrophic events and ask, "why do some accidents turn into catastrophes while others don't?" Two books and thousands of catastrophic events later, Perrow's NAT remains the best explanation yet as to why disasters happen. His seminal work, *Normal Accidents: Living with High Risk Technologies*, is the definitive source for understanding the connection between man-made mistakes and catastrophic consequences. Most of his examples center on critical infrastructure failures that have occurred in nuclear power plants, collapsed energy and power networks, and transportation disasters.

Perrow began his work soon after the 1979 Three Mile Island nuclear power plant reactor partially melted down due to lost coolant and the Bhopal Gas Tragedy—considered the world's worst industrial disaster—on the night of December

2, 1984, at the Union Carbide India Limited pesticide plant in Bhopal, India. "A little reflection on Bhopal led me to invent 'Normal Accident Theory' in order to see how this tragedy was possible" [1]. Perrow's invention provides the basis for understanding all sorts of disastrous events.

Perrow recognized that accidents and minor incidents happen all the time. Most accidents are small and soon forgotten. But some accidents propagate through a *system* of interdependent components and magnify in intensity or severity as they spread, ultimately bringing down the entire system. NAT is a kind of "domino theory," but with the addition of one important element, the size of the dominos increases as more fall.

Perrow realized that linkages among parts of a system are more important than the individual parts themselves. Links are the key to understanding Perrow's theory. "Two or more failures, none of them devastating in isolation, come together in unexpected ways and defeat safety devices—the definition of a 'normal accident' or system accident. If the system is also *tightly coupled*, these failures can cascade faster than any safety device or operator can cope with them, or they can even be incomprehensible to those responsible for doing the coping. If the accident brings down a significant part of the system, and the system has *catastrophic potential*, we will have a catastrophe. That, in brief, is Normal Accident Theory."

The Three Mile Island (TMI-2) nuclear reactor accident in 1979 motivated his two-decade pursuit of NAT. On the surface of it, the TMI-2 accident was highly improbable and completely unexpected. It started when about a cup of water leaked out of the secondary cooling system, which increased moisture in the instrumentation, which in turn interrupted air pressure, which in turn told two pumps to stop, leading to a wind down of a turbine, which caused a buildup of heat that tripped cooling water to release, streaming cold water into a blocked pipe and finally misleading an operator into making the wrong decision. This highly unlikely sequence of events spread and escalated failure throughout the system as temperature continued to rise and the reactor melted down.

TMI-2 happened because of at least five small events (two or more failures) that could easily have been rectified, but were not. They were normal accidents—accidents that one would expect to occur during normal operation. And yet, they spread and magnified the consequences as the nuclear reactor system degraded more with each incident (tightly coupled). This catastrophe is like so many others that start out as insignificant accidents and end up as a "big one" (catastrophe potential). TMI-2 illustrated the principles of NAT in action, just as the Fukushima Daiichi accident illustrated the spread of failure through a series of mishaps, leading to complete devastation.

Perrow devoted a decade studying the big ones like the release of deadly gas from the Union Carbide Bhopal, India, plant that injured over 200,000 people and killed 4,000 in 1984; the Chernobyl, Kiev, Ukraine nuclear power plant fire and radioactive release that exposed 600,000 people and caused the evacuation of 336,000 people; and the space shuttle Challenger disaster in 1986. Indeed, when all known nuclear power plant failures between 1957 and 2011 are tallied and graphed as a consequence exceedence probability, nuclear power is shown to be a high-risk hazard (see Fig. 3.1).

How is *NAT* different than any other theory of accidents? According the Perrow, the difference between an *incident* and an *accident* is the difference between failures of a single component he called a *part* or *unit* and the failure of an entire system. Incidents are bad things that happen to parts of a system. Accidents are bad things that collapse the entire system. NAT distinguishes mundane component failures from dramatic system-wide failures. According to Perrow, coupling among parts and units of a system causes normal accidents. This is what makes a complex CIKR system vulnerable to catastrophic collapse.

Perrow divided systems into four levels of aggregation: units, parts, subsystems, and complete systems. Simple damage to units and parts is called *incidents*, and damage to subsystems and complete systems is called *accidents*. System accidents involve the unexpected interaction of multiple failures. Unexpected and linked failures magnify consequence as they spread throughout the units, parts, and subsystems, eventually collapsing the complete system. Such is the behavior of complex systems and therefore the reason they are considered complex [2].

NAT combines *complexity theory* and the lopsided randomness of power law phenomena into one grand theory of catastrophes—decades before complexity theory was in our lexicon. NAT was the first attempt to understand the physics of catastrophes, and it still works today. The Fukushima Daiichi power plant disaster that devastated Japan 30 years later is a repeat performance of NAT.

In summary NAT is distinguished by three key ingredients: (1) occasionally two or more failures come together in an unexpected way, (2) failures cascade faster if the system is tightly coupled, and (3) when the system has *catastrophic potential*, consequences are also catastrophic. The first ingredient is another way of expressing the power law—the law that describes lopsided unpredictability. The second is new, because it introduces the idea of coupling or connectedness, and the third is an elegant way of expressing *SOC* described below. TMI-2 would not have melted down had it not contained SOC.

3.2 BLOCKS AND SPRINGS

Coupling or connecting one part of a system to another establishes pathways for a fault to spread throughout a system. In complex systems theory, coupling is called *connectivity* or linkage. Such systems are *networks*,

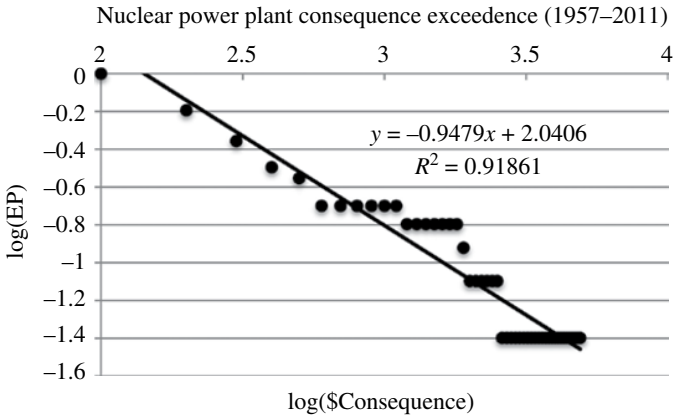


FIGURE 3.1 Consequence exceedence for all known nuclear power accidents (1957–2011) indicates they are high-risk hazards, because fractal dimension is less than 1. Data source: https://en.wikipedia.org/wiki/Nuclear_and_radiation_accidents.

consisting of units, parts, or subsystems called *nodes*, and pathways between pairs of nodes are called *links*. Pairs of nodes are connected by a single link, but the overall wiring diagram of a network has an architecture that determines its dynamical behavior and resilience. In fact, the resilience of a networked system is directly related to this wiring diagram.

Consider the simple apparatus in Figure 3.2. Six blocks (nodes) are connected by springs (links), which allow the blocks to slide back and forth when pushed. Huang and Turcotte used a similar "slipstick" model to explain the tectonic shifting between plates that cause earthquakes [3]. Sliding blocks under the influence of friction and tension is an accepted model of earthquakes, so it should be no surprise that analogous behavior is observed in the computer simulation described here.

Imagine applying an impulse to the leftmost block to force it to move to the right. If the friction between block and plane is small, the impulse will push the leftmost block far enough to collide with the adjacent block. If the force is even stronger, several blocks will be displaced and collide with adjacent blocks. Applying a stochastic force produces an unpredictable number of collisions between adjacent blocks. Therefore, the number of blocks involved in an "accidental collision" is also unpredictable.

Each time the leftmost block is forced to move to the right, it collides with an unexpected number of blocks, depending on the surface friction and force applied. How many blocks does the stochastic force move? This dynamical system is so simple that any sophomore in an engineering curriculum should be able to calculate how many sliding blocks will shift each time the leftmost block is forced to move. But the calculation is not so simple, because as simple as this system appears, its behavior is complex.

Figure 3.2b shows the exceedence probability versus number of moved blocks (as a percentage). Clearly, the exceedence probability of number of blocks moved obeys a power law. Why? The sliding block exceedence distribution is an *extreme statistic*, which has no average value or standard deviation. There is no "typical" behavior. For example, distributions produced by multiplying k random numbers together, recording only the maximum or minimum of a handful of random numbers, and other combinations of random numbers are typically long-tailed as shown in Figure 3.2b. The probability distribution of k blocks colliding in Figure 3.2a was obtained by multiplying random numbers representing the movement of each block. The product of these random numbers forms a distribution that approximates a power law.

The Gutenberg–Richter scale for relating the number of earthquakes of size M is a power law much like the one created artificially by this computer simulation—a program called *SlidingBlocks.jar*. The frequency distribution of sliding blocks can be made to match the Gutenberg–Richter law by careful calibration of the number of blocks, surface friction, and magnitude of forces applied to the leftmost block. There is a logarithmic relation between these factors, so the resulting distribution has a fractal dimension as described in Chapter 2.

The sliding block experiment illustrates Perrow's NAT: The blocks, springs, and plane are units and parts of a larger system. They are connected together by links—the springs and the friction between block and plane—and they slide an unpredictable distance when moved. Their behavior depends on a random accident (a stochastic force), coupling (the springs and friction), and the combined actions of all blocks and springs (catastrophic potential). The sliding block experiment illustrates connectivity, coupling, and stochastic nonlinearity found in most normal accidents. These are also the elements of complex behavior we find in most critical infrastructure systems studied in this book.

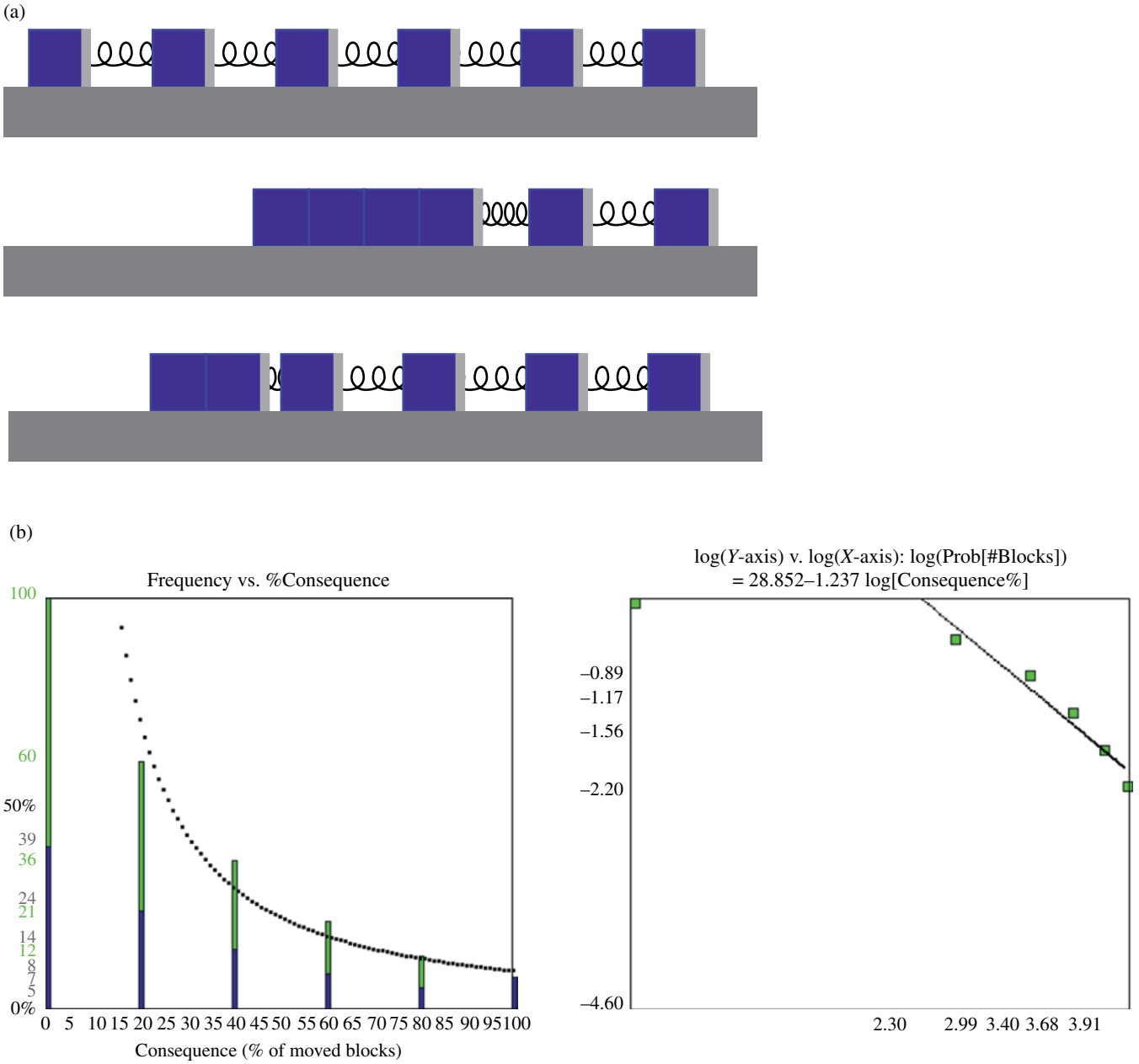


FIGURE 3.2 The six-block apparatus connected by springs illustrates how a simple system can also be complex, because of its behavior. (a) Sliding blocks contract and spring back after an impulse on the leftmost block. Three incidents are shown here. (b) The distance moved by the leftmost block varies randomly, but the exceedence probability distribution is a power law instead of a normal distribution.

3.3 BAK'S PUNCTUATED EQUILIBRIUM THEORY

In the late 1980s Per Bak, Chao Tang, and Kurt Wiesenfeld (BTW) imagined a sand pile formed by dropping grains of sand onto a flat surface and observing the pattern of landslides that inevitably occurred. As the grains of sand accumulate, the angle of repose of the cone-shaped pile increases until reaching a tipping point. Sections of the sand pile would then cascade down the side of the cone, and then the process of buildup and collapse would repeat. The BTW experiment soon became a metaphor for complex systems that appear to be simple but, in fact, behave in unexpected and complex ways.

Figure 3.3a shows several instances of landslides in a simulated sand pile with 45 grains of sand. Figure 3.3b shows the exceedence probability distribution when resistance (friction) is 70%. This means each grain has a 30% chance of sliding when impacted by a grain above it. Bak's sand pile is a lot like the sliding block model of tectonic

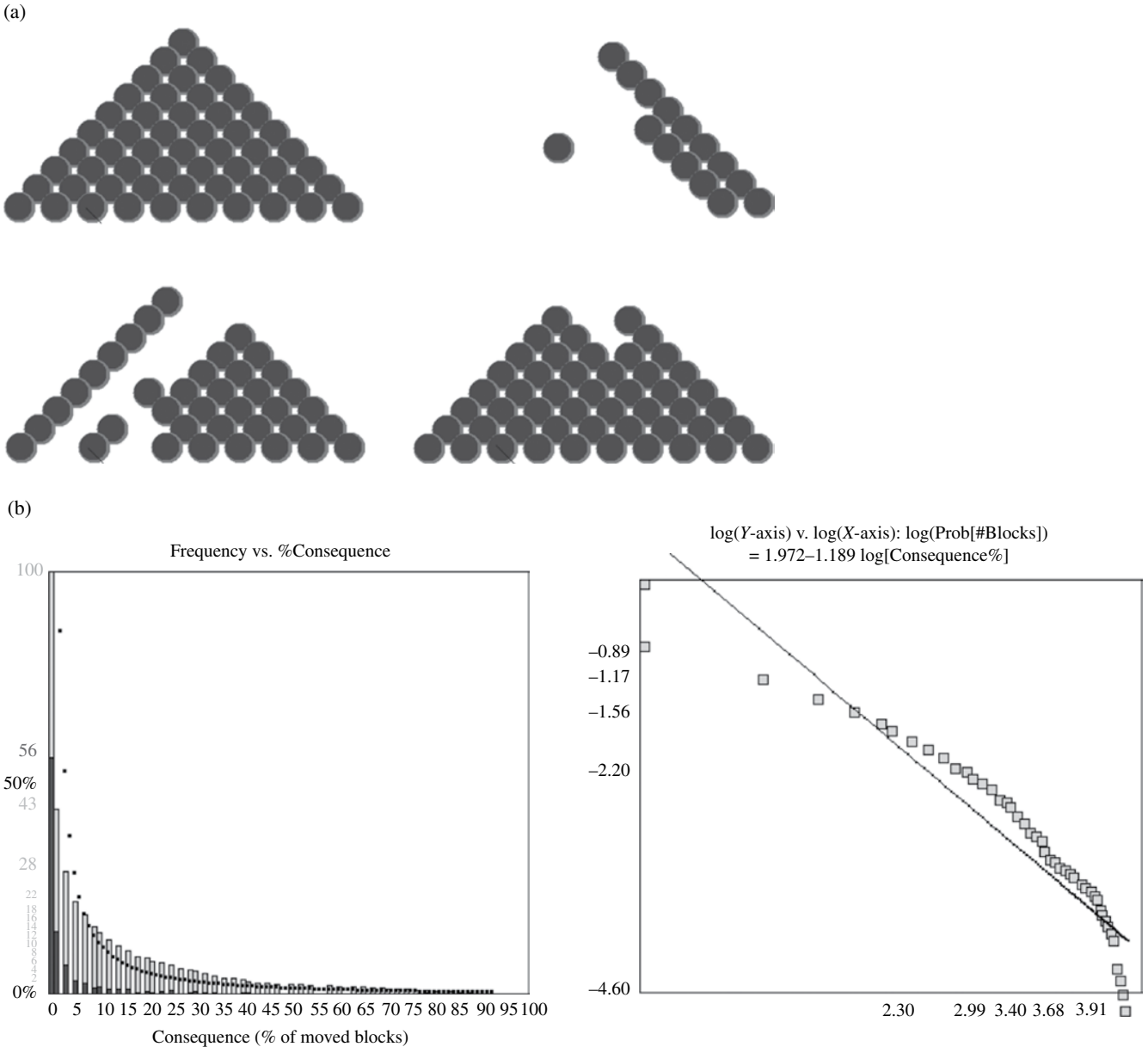


FIGURE 3.3 Bak's sand pile experiment simulates a landslide, but it has become the metaphor for complex system collapse and punctuated equilibrium. (a) Four Bak sand pile landslides illustrate the unpredictable behavior of the sand pile system. (b) The exceedence probability versus percentage of grains of sand collapsed per landslide shows a long-tailed distribution with fractal dimension of 1.97. But the fit to a power law is not perfect.

movement, but in two dimensions. A grain above can start a chain reaction of sliding grains with probability of 30%. In addition, the size of the landslide magnifies as grains join in the action. Bak's sand pile simulates Perrow's mysterious tight coupling and catastrophe potential.

Figure 3.3b shows both frequency and exceedence versus number of sliding grains—expressed as a percentage of the total. Note that both frequency and exceedence are long-tailed distributions. But the log–log plot is not perfect straight line. This imperfection will be explored in Chapter 4, but the reason for it has to do with how self-organization with a complex networked system interacts with the probability of each grain sliding. Briefly, the shape of the exceedence curve is determined by both probability of individual grain movement and the architecture of the sand pile. These two factors will be defined in Chapter 4 as component vulnerability γ and spectral radius ρ.

CIKR sectors are not sand piles, but they have structure similar to sand piles. This structure—defined by the architecture of a power grid, water system, transportation

system, or communication system—determines resilience of the system, as does the threat and vulnerability of the individual components of the system. In fact, this is a fundamental concept of resilient infrastructure stated here as a maxim.

> ***Structure Matters****: The resilience of an infrastructure system is determined by threat, vulnerability, consequence, and the architecture of the system expressed as a network of interacting parts.*

Bak's sand pile is an excellent metaphor for complex systems on the verge of collapse, because the three researchers at the Brookhaven National Labs were *unable* to predict when the landslides would happen or their size—just like normal accidents. The *BTW experiment*, as it soon became known as, is a fitting metaphor for CIP, because critical infrastructure system collapses are unpredictable, too. In addition, it explained Perrow's catastrophic potential in terms of complexity theory. The mysterious catastrophic potential turned out to be SOC—a property of complexity thoroughly studied by Per Bak until his death in 2004.

A fundamental principle of complex systems is a tendency to build up a kind of instability called *self-organized criticality*. Complex systems like the electric power grid, water networks, transportation networks, and communications networks tend to self-organize into a *critical state*, and, once in this state, any change to the system can start a chain reaction. These chain reactions manifest as cascading avalanches, nuclear power plant meltdowns, and electrical power grid collapses. When a sand pile reaches a critical state, the addition of a single grain of sand may lead to avalanches of unpredictable size—even extreme avalanches that completely destroy the sand pile.

Self-organization is one of several fundamental drivers of spectacular failure as demonstrated by power grid outages and power plant meltdowns. The three scientists had independently discovered Perrow's *catastrophic potential*—it is self-organization taken to an extreme level. SOC is Perrow's mysterious force that leads to ruin. As coupling and interdependencies among units and parts of a system evolve, they form an *architecture*—the wiring diagram, if you will—that magnifies system collapse. Most CIKR systems studied in this book suffer from a buildup of SOC due to a variety of factors. For example, California is notorious for its vulnerable energy and power infrastructure. SOC has taken over in the form of deregulation policies, energy marketplace dynamics, tight fuel supplies, utility company financial weakness, growth in consumer demand, lack of generation and transmission capacity, aging infrastructure, and not-in-my-backyard (NIMBY) sentiment [4]. These forces shape the system's architecture, melding it into a metaphorical sand pile on the verge of collapse.

Bak took this idea several steps further: his theory of self-organization became the basis of his theory of *punctuated equilibrium* [5]. SOC leads to bursty behavior and Levy flights as demonstrated in Chapter 2. Bak observed that earthquakes and congestion in communication systems alternate between long periods of relative calm followed by bursts of activity. Similarly, disastrous events occur after long periods of calm followed by bursts of catastrophes, reverting to periods of calm, and so on. The long-tailed power law formed by measuring the elapsed time between events confirms punctuated equilibrium.

Punctuated reality is a feedback mechanism containing two feedback loops. A *normal accident* loop is what we experience most of the time. For example, the Exxon Valdez oil spill in 1989 and the Oklahoma City bombing of 1995 were normal accidents. While they were horrific, they were not equivalent to the 9/11 terrorist attacks or the Chernobyl or Fukushima Daiichi power plant meltdowns in 1986 and 2011. These accidents produce a relatively modest response—oil tankers are now required to be double-hulled, and terrorists receive the death penalty under the Antiterrorism Act of 1996. Unfortunately, reaction to these accidents often contribute to an increase in SOC by increasing the complexity of rules, increasing efficiencies, and optimizing, hardening, and ratcheting up SOC. These reactions start the cycle of self-organization over again, leading to SOC, which contributes to the next catastrophe.

A second feedback loop is more serious. I call this the *black swan* loop, after Taleb's characterization of black swans as rare, extreme, and unpredictable outliers. Black swan events are responsible for extinctions or large adaptations in nature [6]. The terrorist attacks of 9/11 and the financial meltdown of 2008 are examples. They are in a class of their own mainly because of their extremely low probability and extremely high consequences. Their exceedence probability curves are very long-tailed, and the associated fractal dimension is much less than one. And most significantly, they are followed by aftershocks—both literally and figuratively. Aftershocks of this size are typically followed by chaotic adaptation. For example, the counteroffensive against terrorism in the United States, Iraq, and Afghanistan following 9/11 was perhaps as consequential as the terrorist attacks that precipitated the aftershocks in the first place.[1]

Lewis writes, "Black swans are capable of wiping out entire species. For example, the Lake Tobu volcano nearly wiped out the entire human race some 74,000 years ago. Archeologists claim that fewer than 15,000 humans survived this catastrophe, and according to the prevailing archeological theory we are all descendants of roughly 1,000 surviving

[1]Estimates vary, but the United States spent a minimum of $1 trillion on the global war on terrorism, passed the Patriot Act, and established the $60 billion/year Department of Homeland Security as a direct result of the 9/11 attacks.

females capable of reproduction. The Lake Tobu 'volcanic winter' sent the few remaining humans northward out of Africa into the Middle East and Europe. Humanity barely avoided extinction, but we adapted, and reappeared as a mutated and improved species. In fact, human intellect and capability exploded subsequent to this near-extinction event, known as a genetic *bottleneck*" [2].

The two feedback loops are related. The normal accident loop continuously adjusts SOC through incremental optimizations. Machines are made more efficient, computers are cheaper and faster, and more regulations and laws are passed to take into consideration more subtle variations. Energy systems are constantly optimized to squeeze more efficiency from them, and financial systems are optimized to squeeze out more profit. Hospitals eliminate spare beds to save money, and surge capacity is eliminated. Transportation systems are optimized to make them perform better at lower cost.

Each pass through the normal accident loop increases SOC and brings a system closer to the edge of disaster. Eventually the optimized system reaches its critical point so that, when a relatively insignificant event occurs, its effect is magnified. Oil exploration is improved so that drilling 5000 ft under the Gulf of Mexico is not only economical but also more efficient and profitable than ever before. Perhaps corners are cut in terms of safety, or the technology is pushed to its limits. Suddenly, when a small explosion occurs on an ordinary oilrig, the entire Gulf is flooded with millions of barrels of oil. Electric power grids run near their limit, and suddenly, when something insignificant happens in Ohio, the lights go out in New York. When mortgage-backed securities are optimized through packaging and repackaging of derivatives, the failure of an insignificant loan company in Southern California trips a firestorm of financial failures. When intelligence and law enforcement agencies in the United States inadvertently fail to prevent penetration of airport security in Boston, the Twin Towers and Pentagon are successfully attacked, triggering a series of consequences that lead to global warfare and chaotic adaptation around the globe; an unknown virus in China ignites a global disease; and so on. Ever-expanding SOC can be found in diverse systems whether they are physical, biological, virtual, political, or economic.

Bak's punctuated equilibrium model of the world suggests a paradox. Our natural inclination is to optimize and improve efficiency in every possible modern system. In fact, this benefits society by delivering goods and services to large populations at the lowest prices. It not only spurs development and efficient use of resources, but it also increases SOC. Optimization leads to criticality, and criticality leads to catastrophe. Optimized complex systems benefit humanity but contain the seeds of disaster. The longer we postpone the inevitable collapse, the bigger it is. This is *Bak's paradox* and a dilemma for CIP.

3.4 TRAGEDY OF THE COMMONS (TOC)

There is more to complex systems and critical infrastructure networks than risk and SOC. For any CIKR to survive for long periods of time, it must be *sustainable*. Sustainability has many faces, but fundamentally, a CIKR system is sustainable if it remains stable for long periods of time. For example, a drinking water system is sustainable if its source of water is perpetual and it is maintained. An automobile will run forever if it is taken care of and replacement parts are available forever. Of course, infinite sustainability may not be possible, but an infrastructure system must be able to provide a service or commodity for extremely long periods of time to be considered sustainable.

The metaphor for sustainability was provided long ago in the form of the TOC parable [7]. William Forster Lloyd (1795–1852) used a metaphorical pastureland shared by cattle owners to debunk Adam Smith's theory of the "invisible hand." Smith claimed that a hidden law of supply and demand acted as an invisible hand, sorting out imbalances in the economy, without intervention. On the contrary, Lloyd postulated a medieval field of grass (the "commons") shared by cattle owners as a metaphor for the economy. Cattle owners share in the cost of the pastureland but benefit personally when they sell their herd. Acting in their own interest, the cattle owners are motivated to add more and more cattle until the commons is depleted because of overgrazing. By acting in their own self-interest, the cattle owners destroy themselves. This is the TOC.

Instead of an infinitely expandable economy where imbalances are automatically rectified by an invisible hand, Lloyd envisioned a finite resource with limits. Selfish self-interest not only hurt others, but it hurt the predators too. Lloyd's metaphor introduced a limit to infinite expansion called a carrying capacity. If carrying capacity is exceeded, the response is nonlinear and often leads to system collapse.

Garrett Hardin (1915–2003) revised the parable and rejuvenated interest in the concept in 1968 [8]. Hardin pointed out that maximization of profit leads to extinction of both the commons and the cattle. Self-interest leads to self-destruction, which is a kind of paradox, because the purpose of sharing the commons is to improve the sustainability of the community. But if the carrying capacity of the commons is exceeded, the entire system may collapse.

TOC can be used as a test of sustainability of a critical infrastructure system. Some examples—both positive and negative—are as follows:

Water and air pollution damage to crops, buildings, public health.
"Too big to fail" banking that shifts losses to taxpayers.
Over fishing of oceanic fisheries.
Risk transfers in healthcare, maintenance of roads, and so on.

Network effects: individuals buying a cell phone benefits everyone.
Quarantining infectious diseases benefits everyone.
Keeping up the neighborhood increases all home values.

In its simplest form, collapse is caused by depletion of a shared resource by people or organizations acting independently and rationally according to their own self-interest. TOC can diminish the common resource that the person or organization depends on. Thus, individual action may be harmful to the individual or group's long-term best interests.[2]

TOC is a factor in risk and fragility of CIKR, because most infrastructure systems are shared. They form a commons that is hopefully sustainable. But many critical infrastructure systems are not sustainable under current constraints. How do we know if an infrastructure system is sustainable or not? What is the test? A simple simulation of the TOC parable illustrates how to apply *state space* diagramming technology to answer questions of stability and sustainability.

3.4.1 The State Space Diagram

Consider the results of three simulations of the TOC parable shown in Figure 3.4. The first pair of diagrams shows how the balance of cattle and grass in the commons stabilize and reach a stable and sustainable state. The top diagram plots number of cattle and amount of grass over time. The bottom diagram plots the number of cattle *versus* amount of grass. The cattle-versus-grass plot is called a state space diagram and represents the state of the commons system, while the cattle-and-grass-versus-time plot represents the time-varying dynamics of the system. The state space diagram forms a half-circle and then stops at a *fixed point* corresponding to the stable balance between cattle and grass. Once this fixed point is reached, the dynamical system remains there, forever, or until conditions change.

The state space diagram will prove to be more useful in critical infrastructure analysis, because time-varying systems like the TOC commons can be analyzed to determine if they are stable or chaotic (unstable). A system that reaches a fixed point is stable, because the system stays in the state defined by the fixed point. A stable system is a sustainable system.

The second pair of graphs in Figure 3.4 illustrates another possible stable state—sometimes called a *metastable state* because the number of cattle and amount of grass oscillate as shown. Stable oscillations like this show up in the state space diagram as elliptical circuits. An oscillating system never reaches a fixed point, but it is stable, and therefore sustainable, because the commons is not depleted. Instead, the number of cows and grass achieves a dynamic balance that changes over time, never reaches a fixed point, but also never ends.

If the number of cattle and amount of grass get out of balance, as in the third pair of graphs in Figure 3.4, the system becomes unstable and chaotic. Eventually the chaotic oscillations die out and both cattle and grass are depleted. Chaotic collapse shows up in the state space diagram as a line that runs off the diagram. In this case, the state space diagram forms an elliptical circuit and then goes to zero—where it remains forever. Therefore, the state space diagram reveals an instability in the system that leads to eventual collapse. Under these conditions, the commons is not sustainable. Similarly, complex CIKR that are unstable are also unsustainable. While it may take decades for them to collapse, their instability eventually causes their demise.

Figure 3.5 summarizes the meaning of a state space diagram when applied to any dynamical system. The technique plots one influence, a *predator*, against another influence, a *prey*. The cattle owner is considered a predator and the grazing commons the prey in the TOC parable. The state space diagrams of Figure 3.4, for example, plot predator cattle owners against prey grass. Sustainability is achieved when the state space diagram reaches a nonzero fixed point, grows, or moves in a circular path without "blowing up" or reaching zero, as illustrated in Figure 3.5.

3.5 THE US ELECTRIC POWER GRID

The US electric power sector provides a dramatic illustration of the utility of state space diagrams. It is well known that a variety of factors such as regulation (1978 PURPA, 1992 EPACT) and NIMBY have contributed to increasing fragility of the electric power grid. There is not enough long-distance transmission capacity to cope with rising population, rising demand, and trend toward the use of renewable resources to generate power. As the nation transforms into an Internet society, automobiles begin using electric motors, and population growth continues to rise, the lack of resilience in the "middle of the grid" continues to worsen. The power grid is at or near its self-organized critical point.

Figure 3.6a shows what has happened to the power grid since the 1960s. The number of new transmission lines being built rose during the 1960s but steadily declined through the 1970s, 1980s, 1990s, and 2000s. The transmission system circled around a fixed point in the late 1970s due to the Yom Kippur War (1973) and ensuing energy crisis and the 1978 PURPA but then plunged downward rapidly during the 1980s and 1990s. The transmission system circled around a second fixed point in the 2000s but at an extremely low level. Lack of transmission capacity is the predominance reason for low reliability and resilience of the grid.

[2]https://en.wikipedia.org/wiki/Tragedy_of_the_commons

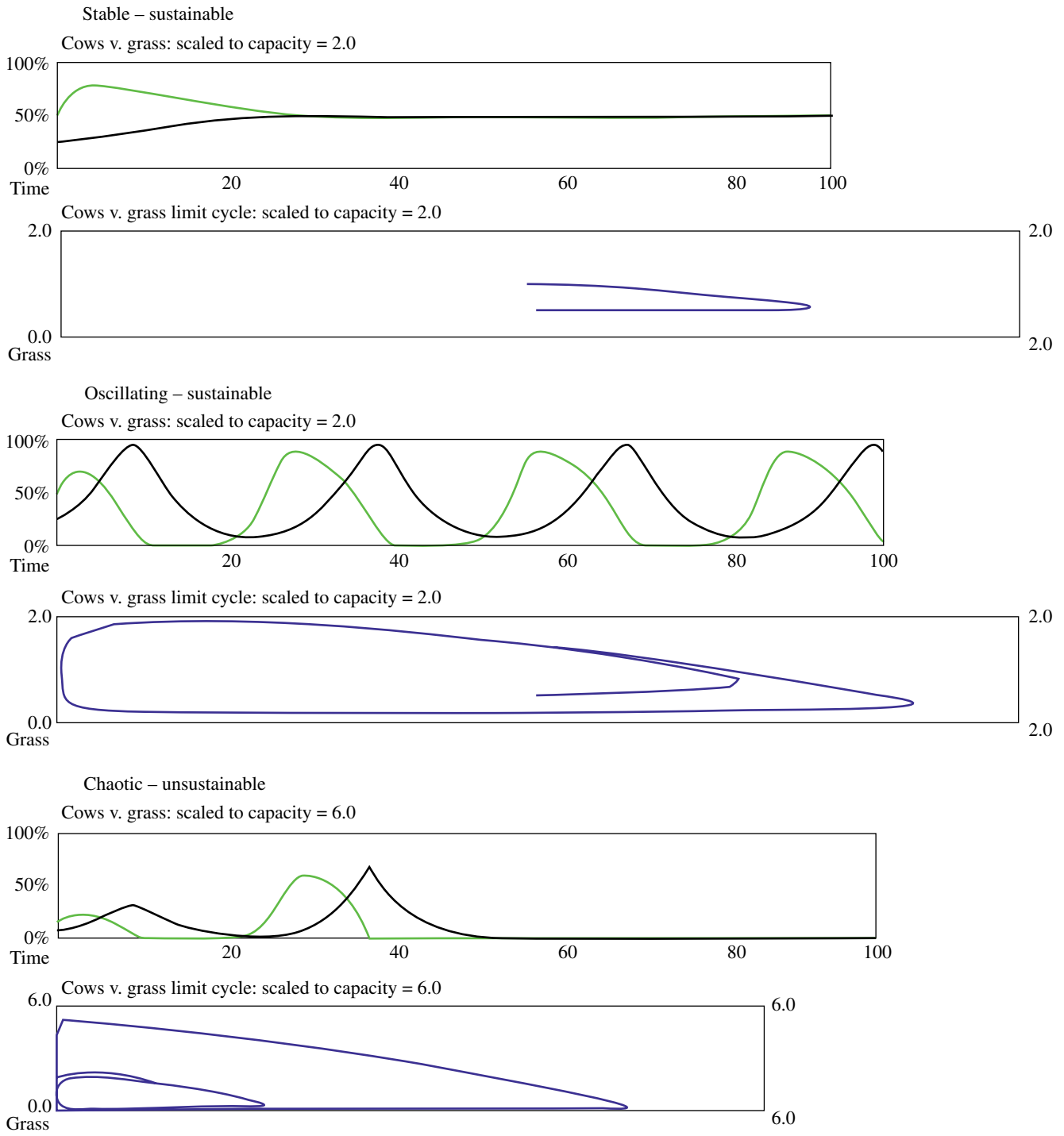


FIGURE 3.4 State space diagrams of three tragedy of the commons scenarios: stable, cyclic, and chaotic.

What is causing the decline in power grid resiliency? Transmission lines are expensive to build, but the nation's gross domestic product (GDP) grew from $2.8 billion in 1960 to $15.5 billion in 2010—over fivefold.[3] How does transmission capacity stack up against GDP? Figure 3.6b contains a second state space diagram of new transmission lines versus US GDP over the same period 1960–2010. Again, the trend is obvious—transmission capacity has not kept pace with economic growth or increasing demand. Transmission line fragility cannot be blamed on the general

[3]At the time of writing, one mile of transmission line costs about $5 million.

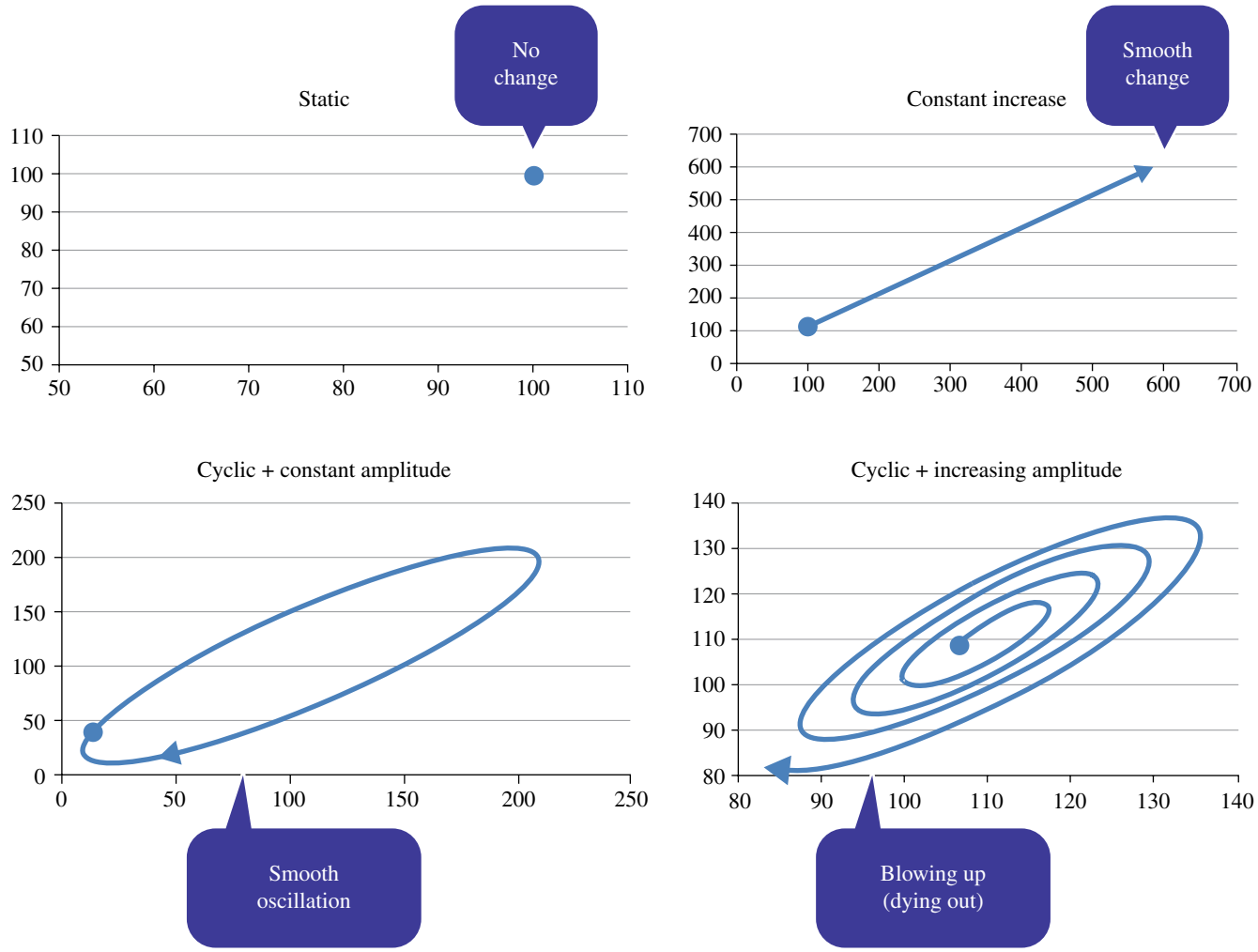


FIGURE 3.5 State space diagrams reveal a system's sustainability or lack of it by plotting predator against prey and observing the resulting patterns.

economy—there must be other factors that are contributing to SOC.

Power grid transmission inadequacy is a classic example of TOC. The tragedy is due to a number of factors, but mainly deregulation policies and energy marketplace dynamics have created an instability in this industrial commons. The 1992 EPACT deregulated electric power markets, which opened transmission to any qualified facility and set prices on what utilities could charge for the use of their transmission lines. As a result, everyone in the industry benefitted from the use of the lines, but nobody was motivated to sustain them. Like the cattle owners in the TOC parable, power companies benefitted from the commons but have no reason to sustain it.

Subtitle D of the massive 2005 EPACT recognized this problem and contains wording to reverse the tragedy by increasing incentives for transmission infrastructure investment:

(a) RULEMAKING REQUIREMENT.—Not later than 1 year after Deadline. the date of enactment of this section, the Commission shall establish, by rule, incentive-based (including performance-based) rate treatments for the transmission of electric energy in interstate commerce by public utilities for the purpose of benefitting consumers by ensuring reliability and reducing the cost of delivered power by reducing transmission congestion.

(b) CONTENTS.—The rule shall—(1) promote reliable and economically efficient transmission and generation of electricity by promoting capital investment in the enlargement, improvement, maintenance, and operation of all facilities for the transmission of electric energy in interstate commerce, regardless of the ownership of the facilities;

(2) provide a return on equity that attracts new investment in transmission facilities (including related transmission technologies);

(3) encourage deployment of transmission technologies and other measures to increase the capacity and efficiency of

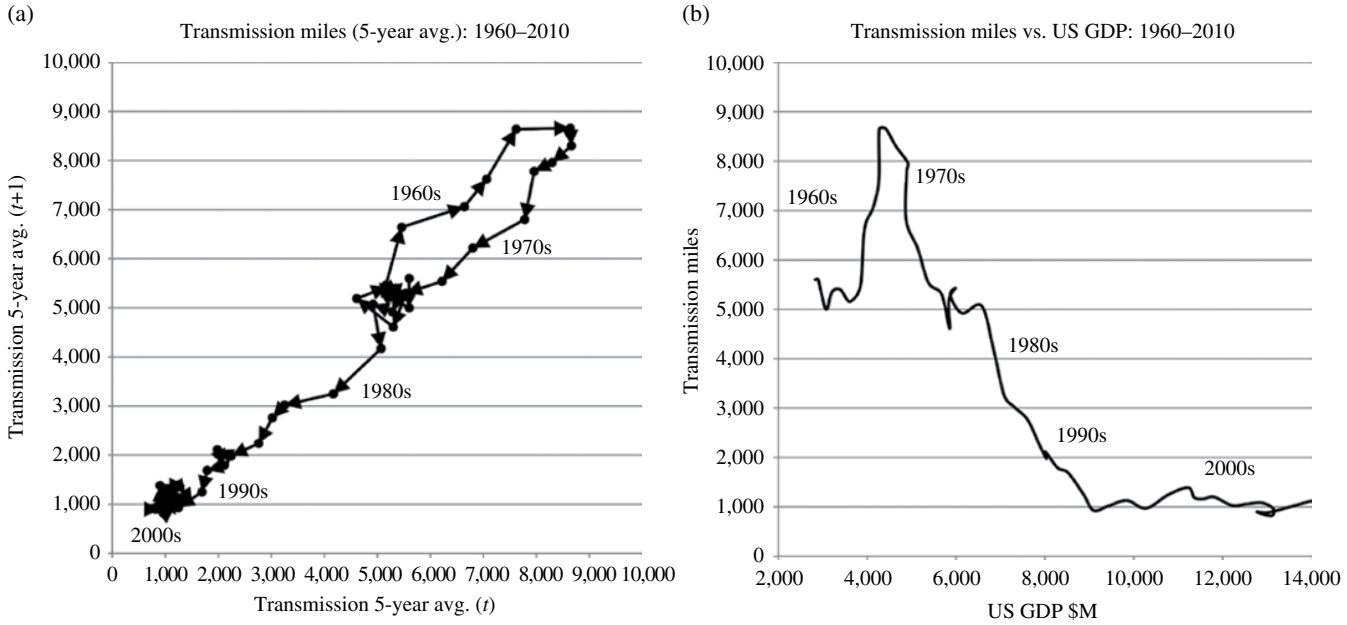


FIGURE 3.6 Two state space diagrams of the electric power grid tragedy of commons show an increasingly fragile power grid due to inadequate transmission. (a) Plot of miles of new transmission lines in year $(t+1)$ versus year t indicates an increasingly fragile system. (b) Plot of new transmission lines in year t versus US GDP indicates the existence of the tragedy of commons in the power grid.

existing transmission facilities and improve the operation of the facilities;

(c) INCENTIVES.—In the rule issued under this section, the Commission shall, to the extent within its jurisdiction, provide for incentives to each transmitting utility or electric utility that joins a Transmission Organization. The Commission shall ensure that any costs recoverable pursuant to this subsection may be recovered by such utility through the transmission rates charged by such utility or through the transmission rates charged by the Transmission Organization that provides transmission service to such utility.[4]

3.6 PARADOX OF ENRICHMENT (POE)

The TOC can ruin a system simply by starving it. But another type of fragility may also exist within a complex system that works in exactly the opposite direction. Instead of starving a system of resources, it is possible to destroy a system by feeding it too much! This is known as the *paradox of enrichment*, and it can ruin a system by supplying it with too much of a resource.

Michael Rosenzweig (1941–), a noted ecologist, discovered a type of instability in the natural world that is often applicable to CIKR systems. Rosenzweig observed, "… If the food supply of a prey such as a rabbit is overabundant, its population will grow unbounded and cause the predator population (such as a lynx) to grow unsustainably large. This may result in a crash in the population of the predators and possibly lead to local eradication or even species extinction."[5] By enriching the food supply (prey) of a predator, it is possible to kill both predator and prey.

POE involves three factors: *predator*, *prey*, and *carrying capacity* of the commons. In the natural world, carrying capacity is the maximum population of a given species that a given environment can sustain. Carrying capacity shows up in physical and economic systems whenever a system degrades due to *enrichment* of the prey. Too much prey feeds an increase in the predator population, which exceeds the capacity of the ecosystem. When this happens, the predator and prey both collapse.

For example, *Braess's paradox* says that adding lanes to a busy freeway (prey) reduces throughput instead of increasing highway capacity. Why? More lanes temporarily expand traffic capacity, but when the expanded volume of cars reaches an intersection, congestion increases exponentially, which produces an even larger delay. Lane capacity leads to more cars, which leads to more congestion, which leads to a slowdown of traffic throughput. For example, doubling the number of lanes of a freeway can multiply the time it takes to travel through one intersection by 127%. In Braess's paradox, carrying capacity is determined by the highway network's intersections, not its roadways.

[4]http://www1.eere.energy.gov/femp/pdfs/epact_2005.pdf

[5]https://en.wikipedia.org/wiki/Paradox_of_enrichment

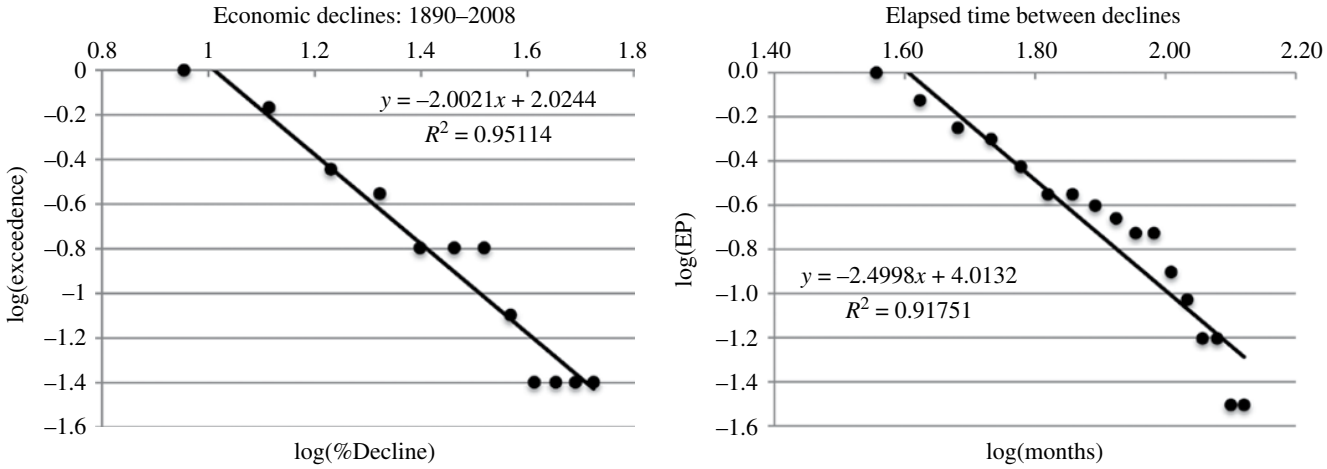


FIGURE 3.7 Economic declines and time to next decline are punctuated events as predicted by Bak's theory.

3.6.1 The Great Recessions

POE offers another explanation of Bak's punctuated equilibrium theory. As a complex system becomes enriched, it eventually reaches and exceeds its carrying capacity. The bubble bursts, and the system returns to a lower state following a period of chaotic adaptation. Then, the process repeats. Excessive enrichment followed by collapse is commonplace in the financial and banking sector—one of the most important critical infrastructures in society.

Financial sector bubbles have been recorded and studied for hundreds of years. The exceedence probability for percentage of decline in industrial production and elapsed time between economic depressions and recessions in the United States suggest they are long-tailed with fractal dimensions in excess of two (see Fig. 3.7). Such a large fractal dimension indicates that most declines are small and most intervals of time between declines are relatively short. Recessions happen often and are relatively brief. Nonetheless, they appear to be inevitable and unpredictable. Why?

The POE in the financial sector goes by a different name—*Minsky moment*. It is a point in time where exuberant and overindebted investors are forced to sell assets to pay back their loans, causing sharp declines in financial markets and jumps in demand for cash [9]. When the financial system creates an excess that exceeds its carrying capacity, the financial ecosystem collapses.

PIMCO's Paul McCulley used the colorful phrase *Minsky moment* to describe the 1998 Russian financial crisis. The Russian economy heated up so much that inflation reached 84% in August, as the exchange rate of the ruble rose from $5.60 to $21.00. The country went into default, banks closed, people lost their life savings, miners went on strike, and protests organized across the country.

Hyman Minsky (1919–1996) first proposed the "The Financial Instability Hypothesis" to explain the vicious cycle that bears his name [10]. Minsky's theory says that in any business cycle investors eventually reach a point where they have cash flow problems due to the spiraling debt incurred by irrationally exuberant speculation. At some point the *price carrying capacity* is exceeded, and there are no buyers for the elevated stock, real estate, bond, or business. Then a major sell-off precipitates a sudden collapse in prices and a sharp drop in market liquidity. Minsky moments are the point at which carrying capacity is exceeded.

3.6.2 Too Much Money

It may seem odd to declare an emergency because too much money is available, but that is exactly what happened in the run-up to the Great Recession of 2008–2009. From 2001 to 2008, the Board of Governors of the Federal Reserve System stimulated the economy to promote a strong economy. Specifically, home mortgage rates were set too low, which enriched a segment of the economy beyond its carrying capacity. This led to a Minsky moment in the housing market (see Fig. 3.8).

The Federal Reserve Act of 1913 gave the Federal Reserve responsibility for setting monetary policy in the financial and banking sector of the United States. The Board of Governors controls the discount rate and reserve requirements—mainly by establishing the amount of money in the system and interest rates on borrowed money. Every day of the business year, money flows in and out of banks because of deposits and withdrawals. Banks must borrow from 1 of the 12 Federal Reserve banks across the county to cover overnight withdrawals and daily deposits. The *federal funds rate* is the interest rate charged to banks on these short-term transactions. The financial structure of the United States is described in greater detail in Chapter 17.

Changes in the federal funds rate trigger a chain of events that affect other short-term interest rates, foreign exchange rates, long-term interest rates, the amount of money and

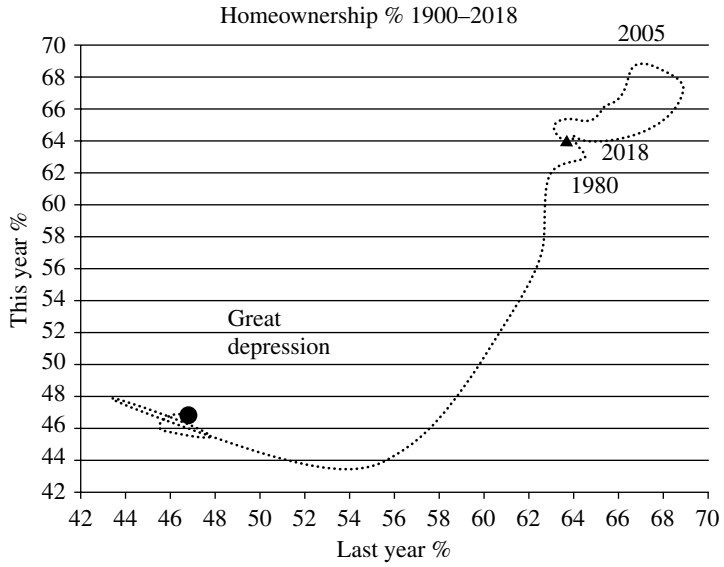


FIGURE 3.8 State space diagram of homeownership from the Great Depression to the Great Recession shows that carrying capacity was exceeded when ownership reached 69%.

credit, and, ultimately, a range of economic variables, including employment, output, and prices of goods and services. Ultimately, a change in the funds rate affects home mortgage rates and in turn the ability for people to purchase a house. When the mortgage rates become extremely low, more people can afford a house. When it dips even lower, many people can afford to buy two or three houses. Cheap money invites speculation, which leads to a Minsky moment.

The Minsky moment in the overheated housing market was reached in late 2008 soon after the market was saturated. Figure 3.8 suggests that the economy at the time was capable of supporting no more than approximately 69% ownership. The state space diagram of Figure 3.8 began to circle a fixed point of approximately 65%. Is homeownership of 65% the carrying capacity of the US economy? This number varies from location to location and across the globe, because the underlying economy varies from location to location.

The financial and banking sector of the United States is regulated by the Federal Open Market Committee (FOMC), which consists of 12 members—7 from the Board of Governors of the Federal Reserve System; the president of the Federal Reserve Bank of New York, the largest reserve bank in the system; and 4 of the other 11 reserve bank presidents: Boston, Philadelphia, Richmond, Cleveland, Chicago, Atlanta, St. Louis, Dallas, Minneapolis, Kansas City, and San Francisco.

Did the FOMC cause the nearly $6 trillion collapse in wealth known of as the 2008 financial meltdown? No, but enrichment did, and the FOMC contributed by enriching the economy. The housing industry was the first segment of the economy to reach its carrying capacity and collapse. This collapse propagated to other segments of the economy, resulting in an economic sand pile landslide.

3.7 COMPETITIVE EXCLUSION PRINCIPLE (CEP)

Critical infrastructure systems are often monopolies or near-monopolies. Consumers want only one water pipe, one power line, one cable TV line, and one telephone line into the house. Competition is slim among highway and railroad owners and operators. There is only one Big Dig in Boston and only one 85-mile long Delaware Aqueduct that supplies most of New York City's water. Infrastructures gravitate toward an exclusive ownership. This leads to one of their major downfalls—lack of redundancy.

Communication and energy infrastructures are particularly vulnerable because of their uniqueness. By 1911 AT&T had become a powerful monopoly—a universally accessible communication infrastructure spanning the nation. The Department of Justice attempted to break the AT&T's monopoly many times over nearly 100 years of on-and-off regulation. But AT&T kept returning as a monopoly or near-monopoly.

It began with a settlement called the *Kingsbury Commitment of 1913*. AT&T divested itself of Western Electric (the equipment manufacturing arm) and agreed to coexist with local independent telephone companies known as local exchange carriers (LECs) today. Western Electric was free to make equipment for AT&T and its competitors. But by 1924 AT&T had bought up 223 of the 234 rivals! AT&T became a monopoly, again.

The Telecommunications Act of 1934 allowed AT&T to continue to operate as a monopoly until the Telecommunications Act of 1996 replaced it. But the Department of Justice brought litigation against AT&T for a full decade, 1974–1984, resulting in dividing AT&T into seven *baby bells*. But once again,

the baby bells merged into three major companies by 2006: Qwest, Verizon, and AT&T.

AT&T was dismantled three times over a period of 130 years, only to return as a monopoly or near-monopoly each time. The company has experienced three major punctuations—the first started in the early 1900s and ended with the 1934 regulation; the second started in the 1930s and ended in 1974–1984 with the baby bells break-up, and the third started in the 1980s and ended in 1996. Is it working on its fourth monopoly? After being dismantled for the third time in the 1990s, AT&T regrouped, and by 2012 it was the eleventh largest company in the United States. And AT&T is a major player in the international Internet infrastructure race—a competition among global corporations to dominate the Internet. In 2012, the "most connected" tier 1 Internet service providers (ISPs) were Cogent/PSI, Level 3 Communications, Inc., and AT&T Services, Inc. Why? The answer is known as Gause's law of competitive exclusion.

3.7.1 Gause's Law

A Russian biologist, Georgii Frantsevich Gause (1910–1986), explained why infrastructure companies like AT&T repeatedly rise to monopolistic positions regardless of obstacles like regulation and competition. Gause discovered what he called the *competitive exclusion principle* in 1932. The principle asserts that in the long run, no two species within an ecological niche can coexist forever. When two species compete, one will be slightly more efficient than the other and will reproduce at a higher rate as a result. The slightly more capable species crowds out the weaker species. The less efficient and weaker species either goes extinct or is marginalized.

The CEP is the biological equivalent of a more general principle known as *increasing returns* in economics, the *network effect* in marketing, and *preferential attachment* among network scientists. These are different terms for the same thing. The idea is simple, but powerful, because it means relatively small advantages can be marshaled into major dominance. Microsoft, AT&T, and other infrastructure companies rose to monopolies by leveraging certain small advantages. Preferential attachment is the fundamental mechanism underlying the architecture of most CIKR, and it the principle means of self-organization in complex systems.

For example, most law enforcement agencies such as a major metropolitan police department operate a *911 Call Center* and information technology department. These systems generally start out small—one operator and one telephone line—and then grow to become a highly centralized but congested IT department. Other major police activities begin to migrate to the fledgling IT/call center. It begins to grow and expand, but it remains centralized because of efficiency and budgetary limitations. Eventually, the IT system becomes a single point of failure. All of the equipment and operators are located in one building, and system response becomes impaired because of limited bandwidth. Gause's law has taken over. The monopolistic IT department becomes the single point of failure—the source of system fragility.

3.7.2 The Self-Organizing Internet

Preferential attachment is perhaps the most common form of self-organization. A complex system evolves from disorder to order while adapting to its environment. In the early years, telephone companies were local, used different technologies, and operated over short distances. The industry was chaotic, not organized. Then AT&T organized it into an efficient, interoperable, long-distance system. Most of the long-distance transmission lines established by AT&T over 50 years ago still exist. They are known as the Long Lines and still form one of the major arterials of the Internet.

There is evidence that self-organization may be taking place in the Internet communication system. The 1996 Telecommunications Act motivates communications companies to colocate in large buildings full of switching equipment known as *telecom hotels*. These buildings contain a variety of voice, data, video, and email switching equipment as well as gateways into the major backbones of the global Internet. This is where many cloud computing systems reside, because it is economically efficient to amortize costs across a number of tenants.

Very large telecom hotels and their tenants form the so-called *tier 1 autonomous system network*, or ASN for short. Every autonomous system (AS) plugged into the Internet is assigned an AS number. Email, video, social media, digital music, and voice all depend on these ASNs. They form the critical pathways through the global Internet. For example, the ten largest ASs known circa 2011 were:

#Links	AS number: AS owner
2972	174: Cogent/PSI
2904	3356: Level 3
2365	7018: AT&T Services, Inc.
1959	6939: Hurricane Electric
1946	701: MCI Communications
1696	9002: ReTN.net Autonomous
1496	3549: Global Crossing Ltd.
1367	209: Qwest Communications
1332	4323: TW Telecom Holdings
1183	1239: Sprint

As you might expect, control of the most highly connected nodes in the ASN goes a long way toward controlling the Internet. In 2011 Cogent/PSI was the largest AS in terms of connections. It is a hub. The largest autonomous AS with the most bandwidth is Deutscher Commercial Internet Exchange, located in Frankfurt, Germany. These ASs obey a long-tailed

distribution in terms of bandwidth and number of connections. The top 20% account for 80% of all Internet traffic.

Preferential attachment is organizing the Internet into a hub-and-spoke architecture where a few super tier-1 ASs dominate the global communications network. Gause's CEP says that an ecosystem has room for only one dominant species. And the species that wins is the one that is able to leverage its relatively small advantage as fast as possible. This means the AS with the most connections and fastest transmission links will dominate all others.

Look at the top 10 highly connected ASs again. AT&T is third in line. It was not even in the top 20 in 2005. The communications ecosystem has expanded to encompass the entire globe. New predators are filling this larger and more expanded ecosystem, each one seeking an edge that will allow it to emerge as the global hub.

The ASN is the nervous system of the twenty-first century. It will run our factories, transportation systems, energy and power systems, food and agricultural production systems, stock markets, and social interactions. Bandwidth will determine the wealth of nations. Note that not a single US company is in the list of the 10 largest bandwidth leaders:

AS owner	Bandwidth (Gb)
Deutscher Commercial Internet Exchange	4029
Amsterdam Internet Exchange	1180
London Internet Exchange	869
Equinix Exchange	946
Moscow Internet Exchange	570
Ukrainian Internet Exchange Network	319
Japan Network Access Point	273
Netnod Internet Exchange in Sweden	204
Spain Internet Exchange	168
Neutral Internet eXchange of the Czech Republic	171

3.7.3 A Monoculture

The Internet also demonstrates another consequence of competitive exclusion. Most Internet servers are either products of Microsoft or various dialects of open source Unix. These two dominant operating systems form a *monoculture*—a single species of software with identical DNA. Malware exploits this commonality across much of the Internet. Every server, desktop, and laptop installed with this monoculture is vulnerable to a contagious virus, worm, and keylogger.

On the other hand, if the Internet ran on a wide variety of operating systems and languages, black-hat attackers would face a much more difficult challenge. They would have to write malware to exploit a variety of flaws in software and protocols for connecting computers together. The monoculture of the Internet simplifies the design of software to exploit weakness in one or two operating systems. But the near-monopoly of Microsoft and Unix software embedded within the Internet increases Internet vulnerability.

The CEP often leads to a monopoly, and the monopoly often leads to a monoculture. Monocultures are easier to attack, because they standardize the interfaces, protocols, and software across an entire system. As the Internet grows and expands to every corner of the globe, its DNA becomes more vulnerable to exploits that could bring the entire Internet down.

3.8 PARADOX OF REDUNDANCY (POR)

Common sense says that a redundant system must always be more adaptable and therefore more resilient. However, when it comes to complex CIKR like the power grid or Internet, redundancy and resilience are often on opposite sides of the equation. Redundancy may in fact make a system less resilient. This is the POR. How can redundancy be bad?

Consider a typical computer system in the IT sector. To increase resiliency against an unexpected hardware failure, the managers of the IT system install a backup computer. For example, the primary computer may be a server that handles all database and Internet processing for a medium-sized police department. The backup computer makes a mirror image copy of every record of data processed by the primary system. If the primary server fails, the backup server takes over immediately. Users never know that a hardware failure happened, because the redundant backup server steps in for the failed primary server.

But what happens if this IT system is attacked by malicious software? The computer virus spreads by copying itself to adjacent computers through their Internet or local area network connections. Thus, the virus spreads to the backup server as well as the primary server. Risk has been increased by redundancy, because expected loss is twice what it would have been with a single server. By making the IT system redundant against a single-point failure, the IT manager has increased risk and decreased resiliency.

System structure in the form of *link percolation* also has a counterintuitive side effect on redundancy. Consider a road network connecting two cities. If all roads from city A connect at one intersection on the edge of city B, adding a redundant road between the two cities increases the congestion on this intersection. Increased congestion reduces the road network's resilience. By increasing the capacity of intercity travel, planners may inadvertently reduce transportation resiliency. This is a type of Braess's paradox.

In the banking sector collapse of 2008, the "too big to fail" theory has a corollary—the "too connected to fail" theory. The financial sector was interconnected in many ways, principally to increase customer base and therefore resiliency. But when a handful of lending agencies failed, the

failure propagated through these business connections to associated businesses. Thus, redundant links once again spread a contagion, and the more connections there are, the faster the contagion spreads. Cascade failure ensued in a dramatic demonstration of a classical normal accident. Unfortunately, redundancy of links increased risk rather than decreased it.

More generally, as redundancy increases by adding more components to a system, the number of connections among redundant parts also increases. As the number of connections increases, the opportunity for greater spreading of cascade failures also increases. SOC increases with density of connections, and so the redundant system becomes less resilient. More connections means more spreading, which leads to a lower fractal dimension, and this equates with lower resilience.

3.9 RESILIENCE OF COMPLEX INFRASTRUCTURE SYSTEMS

The foregoing should give the reader a set of tools for reducing CIKR risk and increasing resiliency. Clearly, risk and resilience is determined by a combination of factors—expected utility and PML risk, SOC, TOC, POE, and CEP. How do these contribute to risk and resiliency in CIKR?

3.9.1 Expected Utility and Risk

PRA and PML risk are overall measures of expected loss and resiliency (fractal dimension). The slope of the exceedence probability curve, plotted on log–log axes—is an overall indicator of the risk and resiliency of a complex CIKR system such as a water, energy, power, transportation, or communication system. A fractal dimension less than one indicates a high-risk system. Therefore, the fundamental objective of CIP is to increase the fractal dimension of threat–asset pairs.

Displacement and time Levy flight exceedence distributions indicate the probability of subsequent events occurring within a certain distance or elapsed time from previous instances. For example, Figure 3.9 shows how the Levy flight data for the billion-dollar weather disasters in the United States and the economic declines over the past century are used to project the likelihood of the next disaster and next economic decline. Of course it is impossible to predict the future. These forecasts are only accurate if you believe the past is able to predict the future.

3.9.2 Countering SOC

SOC is a force multiplier. It increases fragility because of optimizations and efficiencies. In many complex CIKR systems, SOC slowly increases due to economic and regulatory factors. Unfortunately, it takes expensive resources to combat this buildup. Most system owners and operators strike a balance between efficiency and resilience by reducing high concentrations of assets called hubs and eliminating or minimizing bottlenecks called betweeners. Hubs and betweeners can easily become single points of failure.

SOC is a form of *emergence*—the bottom-up process of structuring a complex system by adaptation. The communications sector is a dramatic example. Through a slow but continual process of emergence, the global communications network is evolving toward a hub-and-spoke system much like the commercial airline sector. It is much more efficient to colocate communications equipment—switches, cloud servers, and exchanges—in one building than to spread the equipment throughout the world. Both economics and the 1996 Telecommunications Act drive this adaptation as the global Internet becomes more structured and concentrated.

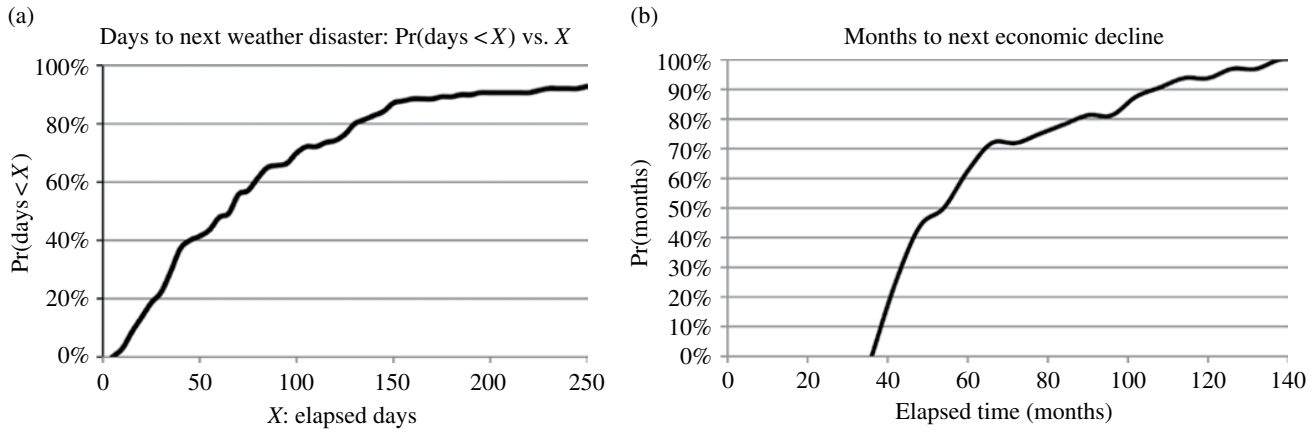


FIGURE 3.9 Probability estimates of the time between subsequent events are based on the Levy flight of each hazard. (a) The probability that the next billion-dollar weather disaster in the United States will happen in X days or less. (b) The probability that the next economic decline in the United States will happen in X months or less.

3.9.3 The TOC Test

The TOC test is used to determine if a CIKR system is sustainable. Any complex CIKR system can be headed for collapse if it is not sustainable. Examples are inadequately funded water and transportation systems and the US power grid under the 1992 EPACT. TOC occurs when the ecosystem of the infrastructure becomes unstable.

Whenever support for a commons is based on an amortized cost-sharing plan, but the products and services provided by the commons are delved out according to self-interest, the commons may become unstable. For example, the Interstate Highway System of the United States may become unsustainable as electric motors replace gasoline motors, because Interstate Highways are supported by gasoline taxes. Public works projects such as drinking water systems may decay and collapse without adequate funding, especially when consumers pay a flat fee to use as much water as they want.

State space diagramming techniques are used to determine if a CIKR commons is sustainable or not. The trajectory of the state space diagram indicates an impending doom, if it heads to zero, an uncontrolled chaos if it heads to infinity, and a metastable state if it circles around. A complex CIKR system may reach a fixed point, where it stays until disturbed by some outside force.

3.9.4 POE and Nonlinearity

POE is one major cause of nonlinearities in complex systems. POE does not always result in disaster, but imminent failure is indicated when the carrying capacity of a CIKR system is exceeded. This can lead to financial bubbles and tipping points. Failure can also occur if the carrying capacity suddenly declines.

A predator–prey state diagram is used to study the effect of POE in a complex system. A state diagram that gravitates toward a fixed point—typically by circling it—usually indicates that the carrying capacity has been reached. Collapses such as a Minsky moment can occur if enrichment pushes the state space diagram too far beyond the carrying capacity.

POE suggests that sudden increases or drops in prey resources can be dangerous. Again, the Interstate Highway System illustrates this concept. The carrying capacity of the gasoline tax base may be exceeded if too many miles of freeway are constructed. By enriching the number of miles of roadway, we may be exceeding the financial carrying capacity of the Highway Trust Fund. The amount of money collected from taxes may become inadequate to support highway maintenance. In fact, the US Congress has had to add billions to the Highway Trust Fund to cover this enrichment of miles. (Additionally, Highway Trust Funds have been diverted to non-highway projects.)

It is easy to see the carrying capacity of many CIKR system decline as infrastructure ages and maintenance outstrips the budgets of states and cities. Obviously, POE is closely related to TOC. What is the difference? TOC collapses occur when self-interest outstrips support of the commons. POE collapses occur due to a nonlinearity caused by adding too many resources to the underlying infrastructure (miles of roadway), which temporarily increases capacity (more cars) but eventually collapses because more predators overshoot (cars increase but use less gasoline). POE introduces an instability due to enrichment of the commons, while TOC may cause collapse because the predator demands too much of the commons.

3.9.5 CEP and Loss of Redundancy

Application of the CEP can lead to loss of redundancy and surge capacity—two of the most significant causes of fragility—because it eliminates competition, diversity, and backup capacity. The energy, power, communications, and public health sectors are challenged by this principle. Energy companies build single pipelines from the Gulf Coast of the United States to markets in the Northeast. Power companies lack surge capacity. Communications companies tend to become monopolies and universal service suffers. Hospitals minimize spare beds, expensive equipment, and operating rooms to save money. Few communities have competing hospitals, which means they have minimal emergency room capacity.

Competitive exclusion is an application of preferential attachment. When one species, say, a hospital, gains an advantage over another species, say, a cancer center or heart center, the other species—other hospitals—must adapt or die out. The advantage of one hospital over another soon gains momentum, which leads to the advantaged hospital gaining more advantage. This feedback increases the distance between first place and second place. Eventually, the dominant species—one hospital—becomes a regional monopoly. The result is less resilience under emergency conditions.

There is no cure for CEP without introducing idle capacity, redundant facilities, backup buildings and equipment, and infrequently used surge capacity. These antidotes and redundancies cost money. Therefore they are often eliminated. Unless a business can justify redundant capacity through dual-use technologies, CEP will reduce resilience.

For example, hospital surge capacity might be increased by the use of public school buildings during an emergency. Fire and police departments might increase redundancy by partnering with neighboring fire and police departments. The US Department of the Interior fights massive forest fires by aggregating fire fighters from adjacent states. Many water and irrigation districts combine resources to fend off droughts and manage floods.

3.9.6 POR and Percolation

The POR goes against common sense, because it says that redundant systems can be less resilient to cascade failures. This is a by-product of complexity—as a system becomes percolated, more nodes and links are added; it becomes more self-organized. As it becomes more self-organized, it becomes more susceptible to cascade failures. Cyber exploitation of the Internet is the most dramatic example.

Redundancy may increase robustness of a system against physical attack but decrease its resiliency against cascade failures. Therefore, policy-makers and strategists must be careful to analyze the risk implications of robustness. Does a robust CIKR system always improve risk? Or does redundancy and robustness increase SOC by increasing concentrations of assets, interdependency, or choke points?

3.10 EMERGENCE

Emergence is how complex systems adapt to their environment through an evolutionary process of bottom-up interactions. Generally, an overall pattern or structure emerges as simple interactions take place at a local level. For example, the massive communications network that currently spans the globe started out as a local phenomenon—local telephone companies connecting homes by relatively short wiring. Then local community telephone companies merge into larger regional companies. AT&T connected only 50% of the US population 68 years after its creation. And now, AT&T and other communications companies are joining their networks together to form the global Internet. Plans already exist to extend these networks to other planets in the solar system.

Emergence of an infrastructure has another characteristic—it evolves systems from relatively chaotic and unstructured beginnings to relatively structured and stable states. An unstable organism cannot survive for very long, so successful CIKR systems evolve toward stability and order. However, the type of order matters. As a CIKR system evolves from a random to a *scale-free* structure, it becomes more vulnerable to collapse.[6] Unfortunately, most critical infrastructure in the world tends to become scale-free—a topic explored in more detail in Chapter 4.

3.10.1 Opposing Forces in Emergent CIKR

While preferential attachment is the main force driving the formation of CIKR systems, it is not the only one. Higher-order forces either accelerate or retard emergence in nonlinear and sometimes unexpected ways. For example, the CEP drives systems toward monocultures and monopolies, but TOC drives them toward creative destruction and extinction. These opposing forces are often balanced by governmental restrictions—typically regulation. In fact, the past 100 years of critical infrastructure evolution has been shaped largely by governmental interference.

CIKR sectors form an industrial commons under the auspices of federal, state, and local government regulators. Figure 3.10 illustrates how the delicate balance between competitive exclusion in an entirely free-market economy and TOC in an entirely regulated command economy has shaped the CIKR studied in this book. These commons contain value chains connecting predators and prey, as shown in Figure 3.10. They also implement different business models, depending on the regulatory world they exist in. Some are near-monopolies, and others are near-socialistic command economy value chains.

At one extreme is the transportation commons formed by the Interstate Freeway System developed by the US federal government and paid for by fuel taxes. In this commons, the predator is the tax-collecting government, and the prey is the traveler who pays through taxes whether he or she uses the freeway or not. The freeway system is subject to a tragedy if consumers use up the freeways faster than taxes can replenish them. On the other hand, privately owned freeways tend to become monopolies because redundant freeways are not profitable.

At the other extreme are deregulated monopolies like the commercial airline commons formed by licensed air routes and landing rights at airports. Commercial carriers prey on the commons and exploit airports and consumers for financial gain. Everyone pays for what they use, and nobody is forced to fly, but the commons is largely controlled by industry. As a result, the airline industry has been consolidating according to Gause's law since deregulation in the 1970s. Will a monopoly emerge, or will a tragedy ruin the commercial airline commons?

In between these two extremes are several CIKR sectors that are partially deregulated and partially regulated through a complex set of rules and price controls. Energy, power, and communication commons are open to competition but under rather restrictive rules of engagement. These rules evolved over the past century to balance the tendency for infrastructure commons to become monopolies versus nationalized industries under command economy control—perhaps paid for by taxes. Under rules established by years of litigation and Congressional action, these commons are partly open to competition and partly closed according to Gause's law.

The commons formed by the electric power value chain illustrates this balance. Electric utilities must open access to their transmission lines under regulated toll pricing, but regulation also prevents their disruption by competitors that

[6]The number of connections at each node of a random network obeys a binomial distribution. The number of connections at each node of a scale-free network obeys a power law. Scale-free means a system has a hub with far more connections than the average and many nodes with far fewer connections than the average.

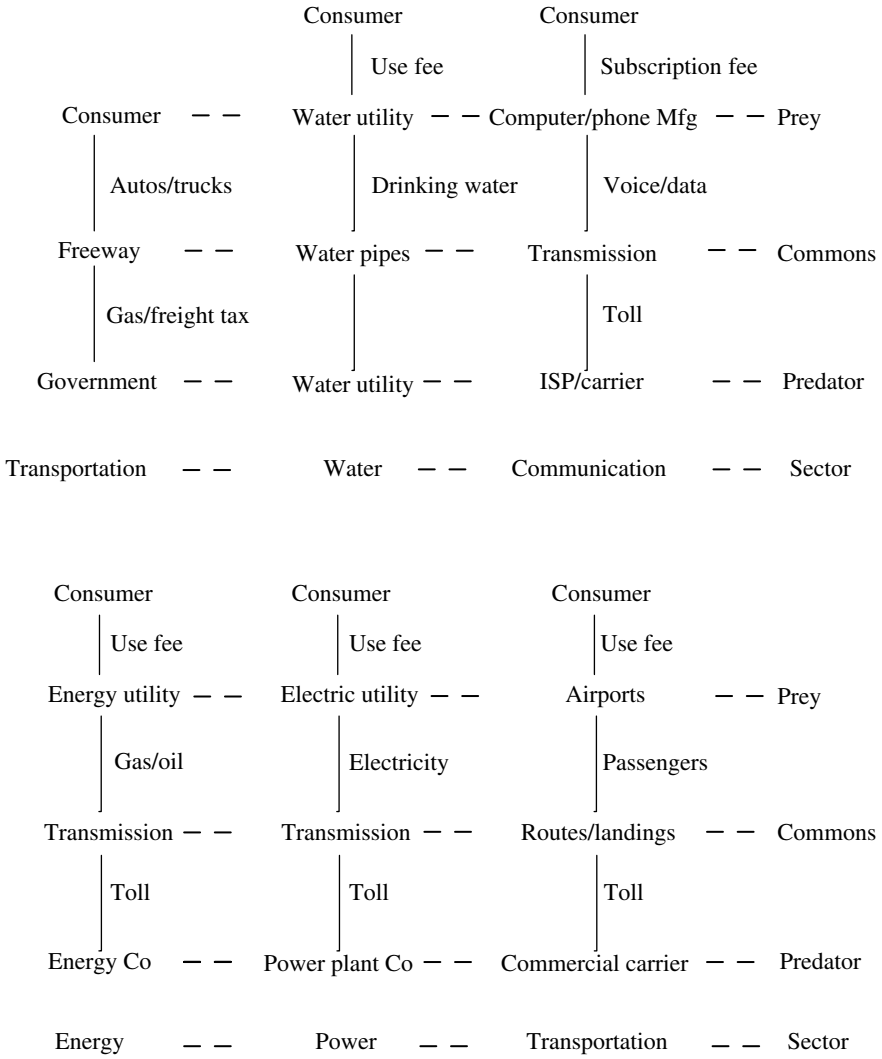


FIGURE 3.10 Most CIKR sectors form industrial commons around value chains as illustrated by several sectors.

might bypass them using new technologies such as distributed solar generation or the creation of consumer-owned and consumer-operated utilities. In most parts of the country, it is impossible for a homeowner to make a profit from his or her backyard solar farm by selling surplus electricity to a utility. There is no EBay.com or Amazon.com of electricity because regulation still protects this commons from free-market competition.

3.11 EXERCISES

1. What are the elements of NAT?
   a. Coupling, two or more faults, catastrophic potential
   b. Tight coupling, links, catastrophic potential
   c. Diversity, complexity, links
   d. SOC, TOC, POE
   e. Risk, exceedence, PML
2. The BTW experiment was also known of as:
   a. Punctuated equilibrium
   b. Black swan
   c. Sand pile metaphor
   d. Power law
   e. Ranked exceedence
3. Punctuated equilibrium explains why:
   a. The 9/11 terrorist attacks happened.
   b. The Gutenberg–Richter law is a power law.
   c. The ranked exceedence probability is a power law.
   d. Normal accidents happen.
   e. Catastrophes are bursty.
4. Efficiency and optimization of systems explains:
   a. SOC
   b. TOC
   c. POE
   d. CEP
   e. PML risk

5. One test of sustainability of a CIKR is:
   a. PRA risk analysis
   b. TOC
   c. Resilience
   d. Redundancy
   e. Surge capacity
6. TOC and POE model CIKR as a system containing:
   a. A commons
   b. Predators and prey
   c. State space
   d. All of the above
   e. None of the above
7. Carrying capacity is a key element of:
   a. Risk
   b. PML risk
   c. CEP
   d. Resilience
   e. POE
8. In economics, POE is called:
   a. Braess's paradox
   b. Minsky moment
   c. Homeownership
   d. PML risk
   e. None of the above
9. Stability and sustainability of a CIKR system can be analyzed using:
   a. PRA
   b. Fixed point analysis
   c. True exceedence probability
   d. Gutenberg–Richter law
   e. State space diagram
10. Preferential attachment is a major:
    a. Theory of catastrophes
    b. Explanation of SOC
    c. Explanation of TOC
    d. Source of resilience
    e. Source of surge capacity
11. Competitive exclusion reduces resilience, because it:
    a. Removes redundancy
    b. Removes surge capacity
    c. Increases SOC
    d. Creates a hub-and-spoke architecture
    e. All of the above
12. Monocultures are one possible consequence of:
    a. CEP
    b. POE
    c. TOC
    d. PML risk
    e. True exceedence
13. A high-risk hazard has:
    a. Large PML risk
    b. Long-tailed exceedence probability
    c. Fractal dimension greater than one
    d. Large consequence
    e. History of black swan events
14. The sliding block or slipstick model of collapsing infrastructure illustrates:
    a. Long-tailed hazards
    b. Complex behavior
    c. Extreme statistics
    d. Distribution of the product or random numbers
    e. All of the above
15. The Gutenberg–Richter law is an example of:
    a. True exceedence probability
    b. Ranked exceedence probability
    c. 10 sliding blocks
    d. BTW experiment
    e. PML risk

3.12 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. Are the five sources of complex system stress described here (TOC, POE, CEP, SOC, POR) independent factors? Explain why they are either separate effects or if they overlap, how they overlap.
B. Why is POR a paradox? How can redundancy be bad?
C. What is the relationship between emergence and scale-free structure and vulnerability to collapse? Is scale-free structure the only form of emergence? Is it the only form of structure that leads to vulnerability to collapse?
D. How does the self-organization of the Internet affect its ability to prevent or reduce the spread of malware? How does the Internet's monoculture affect the spread of malware?
E. How are Perrow's NAT, Bak's punctuated equilibrium, SOC, and resilience related?

REFERENCES

[1] Perrow, C. *Normal Accident Theory*, Princeton: Princeton University Press, 1999, pp. 356–357.

[2] Lewis, T. G. *Bak's Sand Pile*, 2nd ed, Monterey: AgilePress, 2011, pp. 368.

[3] Huang, J. and Turcotte, D. Evidence for Chaotic Fault Interactions in the Seismicity of the San Andreas Fault and Nankai Trough, *Nature*, 348, 6298, November 15, 1990, pp. 234–236.

[4] Rinaldi, S., Peerenboom, J. P., and Kelly, T. K. Identifying, Understanding, and Analyzing Critical Infrastructure Interdependencies, *IEEE Control Systems Magazine*, 21, 6, December 2001, pp. 11–25.

[5] Bak, P., Tang, C., and Weisenfeld, K. Self-Organized Criticality: An Explanation of 1/f Noise, *Physical Review Letters*, 59, December 2001, pp. 381–384.

[6] Taleb, N. N. *The Black Swan: The Impact of the Highly Improbable*, New York: Random House, 2007.

[7] Lewis, T. G. *The Book of Extremes: Why the 21st Century Isn't Like the 20th Century*, Cham: Copernicus Books, 2014.

[8] Hardin, G. The Tragedy of the Commons, *Science*, 162, 1968, pp. 1243–1248.

[9] Lahart, J. In Time of Tumult, Obscure Economist Gains Currency, *The Wall Street Journal*, 2007. Available at http://online.wsj.com/public/article/SB118736585456901047.html. Accessed August 18, 2007.

[10] Minsky, H. P. The Financial Instability Hypothesis, Working Paper No. 74, May 1992, pp. 6–8.

4

COMPLEX CIKR SYSTEMS

A *complex CIKR system* is a collection of components or parts—CIKR assets—that are *interrelated*, *interdependent*, and *linked* through many interconnections and behave as a *unified whole* in *adapting* to changes in the environment.[1] Generally, these systems appear simple on the surface but often turn out to behave in complex ways due to six major environmental factors: *threat*, *efficiency*, *regulation*, *cost*, *NIMBY*, and *demand* or *load* on the system. The dynamic interaction among CIKR assets and these environmental factors is what makes complex CIKR systems complex.

Complexity theory is especially useful for the study of CIKR, because most infrastructure systems appear to be simple structures—roads, bridges, pipelines, power lines, communication links, transportation networks, and so on—but in practice, they have been found to behave like well-known complex systems. They evolve over long periods of change, shifting from initially disordered collections of assets toward greater order and structure as they adapt to efficiency and optimization forces. They often behave in unexpected—long-tailed—manner, and when under stress, they fail like the metaphorical sand pile. While their behavior under stress is unpredictable, they almost always obey a power law, as described in Chapter 3.

Complexity provides a unified theory for understanding and fixing fragile CIKR. In particular, this chapter develops several key metrics that give great insight into what causes fragility and what might be done about it. First, by observation we know that large and complex CIKR systems such as power grid, Internet, or transportation systems evolve toward greater levels of self-organization according to Bak's theory. Self-organized criticality (SOC) is quantified in terms of spectral radius—a measure of network structure. Spectral radius increases with the density of links and size of heavily connected hubs. Resilience against cascade failures also decreases with an increase in spectral radius.

Spectral radius is perhaps the most important property of a CIKR as far as risk and resiliency is concerned, but it is not the only measure. Betweenness—the number of paths running through a node—is a measure of self-organization that is particularly useful for understanding the fragility of flow networks—networks that deliver a flow such as electrons, water, gas and oil, and so on. High betweenness indicates a potential chokepoint and therefore a node or link in the CIKR system that quickly becomes overloaded when the system is stressed.

A third property is useful for analyzing the carrying capacity of system under attack by a cascade-causing force such as an epidemic in human and Internet populations, and rolling collapses in transportation and energy systems are *critical links* and *blocking nodes*. Critical links and blocking nodes are essential to maintaining connectivity of a CIKR system. If a critical link or blocking node is removed, the CIKR system separates into disjoint islands and can no longer work as a unified whole.

A qualitative framework for evaluating whole-of-government complexity as a homeland security enterprise and, without math, reduces the factors contributing to complex responses to catastrophic events rather than attempting to

[1]http://www.businessdictionary.com/definition/complex-adaptive-system-CAS.html

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

model CIKR as a network. This approach has been advocated by Lori Hodges and is described at the end of this chapter. The Hodges Fragility Framework homes in on community fragility—or the lack of it—and identifies four causal relationships: connectedness (a network-like model), stability of the community leadership, and sustainability (recovery after an event). The nonquantitative reader may want to gloss over the network model and concentrate attention of the Hodges model. The sidebar is an alternative, nonmathematical survey of the more rigorous network science ideas in this chapter.

SIDEBAR 4.1

Network science is a rigorous approach to CIKR risk and resilience assessment based on modeling CIKR as networks. The approach yields quantitative results, but it may also go beyond an analyst's requirements for rigor. Instead, a qualitative approach combined with a general understanding of network science may be adequate. This sidebar may be a substitute for the more rigorous and detailed explanation.

Nodes and links may represent the assets of a complex CIKR. Nodes are assets such as computer servers, water processing facilities, buildings, and so on. Links are connectors such as Internet cables, pipes, and roadways. The manner in which nodes and links are connected is called the network's topology and plays a key role in determining risk and resilience under stress such as a cascading electric power network or congestion in a road network. Topology is essentially the "wiring diagram" of a network that defines which nodes are connected to one another through links.

Networks can be broadly classified as random, scale-free, or clustered, depending on their topology. A random network randomly assigns connections between pairs of nodes, so the distribution of connections obeys a binomial distribution. Scale-free networks assign connections between pairs according to preferential attachment. Preferential attachment means connection probability is biased so that a few nodes are highly connected and many nodes are barely connected. The distribution of connections obeys a power law. A clustered network assigns connections such that clusters of nodes are highly connected. Clusters are connected by "long-distance" links.

Scale-free and clustered nodes are considered more structured than random networks. In practice, most CIKR networks are structured, and many are scale-free. The reason this matters is that scale-free networks are more fragile than random networks. The spread of malware in the Internet is exacerbated by the fact that the Internet is a globe-spanning scale-free network. Topology matters, and the more structured a network, the more important topology is to its resilience.

Individual nodes and links are vulnerable to a threat–asset pair as described in previous chapters. A simple TVC product can be used to define static risk, but this is inadequate to describe the dynamics of an entire system represented as a network. Two dynamic behaviors are of interest in this formulation: cascading and flows. One triggers cascade failures or more faults (a failed node or link) followed by a "domino effect" whereby adjacent nodes and links fail with a given probability, much like the spread of a disease through a population. Flows are different—a fault in a flow network may cut off the flow of a commodity such as oil in a pipeline network or automobiles in a road network. Flow risk is measured in terms of the loss of output from a flow network, while cascade risk is measured in terms of the total consequences accrued from faults of multiple assets.

The purpose of network analysis is to analyze a complex CIKR system in terms of risk and resilience and to understand the causal relationships between critical factors and risk and resilience. That is, we want to know what leads to risk and loss of resilience so resources can be allocated to the most critical assets. Clearly, homeland security professionals want to apply limited budgets to the most critical assets in a CIKR in order to optimize the benefits.

The question is, "what is critical about a critical infrastructure system?" The answer lies in applying network science properties to CIKR systems modeled as a network. The criticality factors described here address both types of risk and resilience: loss due to cascading and loss due to disrupted flows. The following criticality factors and their causal relationship with risk and resilience are explained in detail in the body of this chapter, but in summary, they are as follows:

- Vulnerability: Risk goes up with an increase in vulnerability and resilience goes down.
- Connectivity: Cascade risk goes up with an increase in connectivity, and resilience goes down. Targeting of high-connectivity assets exacerbates cascade risk and cascade fragility.
- Influence: Cascade risk goes up with an increase in influence, and cascade resilience goes down. Targeting of high-influence assets exacerbates cascade risk and cascade fragility.
- Betweenness (bottleneck): Flow risk goes up with an increase in betweenness, and flow resilience goes down. Targeting of high-betweenness assets exacerbates flow risk and flow fragility.
- Overloading: Assets may become overloaded when a failure of one asset increases the load on another asset. Thus flow risk is higher for overloaded assets. Flow resilience is defined in terms of overloading ratios—as the ratio of normal

operation versus stressed operation (due to an outage or rerouted flow) increases, resilience decreases. Therefore, flow resilience depends on overload ratios.

- Blocking nodes and links are assets with "infinite" betweenness. A blocking node is critical to both cascading and flows. Hardening a blocking asset so that it does not propagate a fault stops a cascade. Flow resilience can be improved by protecting a blocking asset so that it does not overload and fail.

Tools like MBRA exist for automatically analyzing complex CIKR modeled as a network. These tools typically automate the computations needed to identify criticality factors and assign risk individually and as a dynamic property of cascading and flow networks. Given a limited budget and knowledge of critical factors as described above, a software tool such as MBRA can optimally allocate resources to the most critical nodes and links such that risk is reduced.

Network models connect cause and effect in a precise and unambiguous way. Analysis of a model allows the security analyst to identify the assets most responsible for cascading or loss of flow and then allocate resources for the best return on investment (ROI). Quantitative technologies such as network modeling answer the question of "what is critical," so that an owner/operator of a CIKR system can reduce risk and enhance resilience in the most optimal manner. The value of quantitative modeling, then, is to optimally allocate resources that produce the highest ROI.

The goal of this chapter is to develop a unified theory of complex CIKR and illustrate how critical factors can be identified and used to improve resilience of complex CIKR systems:

- *CIKR as networks*: Most CIKR assets are part of a system represented as a network G = {N, M, F}, where N is a set of nodes, M is a set of links, and F is a mapping function that defines how node pairs are connected. Networks have several critical factors of special interest to the study of CIP: the *connectivity* (aka degree) of a node is the number of links connecting it to other nodes; *betweenness* of a node or link is the number of shortest paths through the node/link on all routes to/from all nodes; *cluster coefficient* is a measure of how connection density among neighboring nodes; height is the number of hops from a node to the furthest source or destination node in the network; and *influence* is a measure of the impact a node has on its neighbors.
- *Critical factors*: A critical factor is a causal property of a node/link that relates cause and effect. For example, vulnerability, connectivity, influence, betweenness, overloading, and blocking properties contribute to cascade failure and cascade resilience or lack of it; bottleneck betweenness, overloading, and blocking nodes/links contribute to loss of flow due to a failed node/link. Controlling these factors is a method of controlling risk and resilience.
- *Types of networks*: Networks can be classified according to how links are distributed to node pairs. There are three fundamental types of networks: *Random networks* are formed by randomly connecting pairs of nodes together. Links are distributed according to a binomial distribution, as if a coin was tossed to determine how nodes are connected to one another. *Scale-free networks* are networks containing a hub with many connections and many nodes with only a few connections. Connectivity of nodes via links is distributed according to a long-tailed power law. *Clustered networks* have no particular link distribution, but instead, their nodes are more tightly connected to one another in local clusters. *Cluster coefficient* is high in a clustered network because the nearest neighbors of a node are connected to each other as well as the node. Clusters of highly connected nodes are typically connected to one another through a sparse number of "long-distance" links. Think of long-distance links as freeway routes connecting cities with their dense fabric of local roads within the city limits.
- *Self-organization (SOC)*: Networks exhibit self-organization when they transition from random to scale-free or clustered networks. Self-organization is a form of emergence that works from the bottom up. Local changes, such as percolation or switching of links, over long periods of time, result in some form of order or structure. Self-organization of CIKR networks manifests as high-degreed hubs and, to a lesser amount, by betweeners and clusters. This is due to emergence, which is typically traced back to the six major environmental factors: *threat*, *efficiency*, *regulation*, *cost*, *NIMBY*, and *demand*. Spectral radius is a measure of self-organization and is equal to the largest influence factor among all nodes.
- *Spectral radius*: A macroscale measure of SOC in networks of all types is the spectral radius. It measures both percolation density and link structure. High spectral radius indicates "more structure" than low spectral radius. The spectral radius of a random network is typically close to the mean degree of the network. (Mean degree is 2(#links)/#nodes).) The spectral radius of a scale-free network is typically many times greater than the mean degree, due to its highly connected hub.
- *Percolation and SOC*: Link percolation is the process of adding links to a network. It reduces the diameter of a network and often produces a *small world* (a large network

with small diameter or small number of hops to get from one node to any other). Percolation increases spectral radius, which is a measure of SOC, and reduces cascade resilience. Highly percolated networks typically contain redundant paths, which increase robustness, but may introduce the *paradox of redundancy (POR)*.

- *Preferential attachment and hubs*: Preferential attachment is the most common type of emergence in a complex CIKR network. It is also the mechanism of the *competitive exclusion principle* discovered by Gause (see Chapter 3). Preferential attachment generally creates a scale-free network with high spectral radius and is generally a by-product of optimization, cost-reduction, and other efficiency measures.
- *Cascading networks*: The most common type of normal accident or system failure in a CIKR is cascading—the domino-like spread of collapse throughout a network. In this book cascading is modeled as a contagion—a single fault propagates from node to node through links with probability defined by *vulnerability*. As the density of links or vulnerability increases, so does the spread of a cascade.
- *Cascading resilience*: A network's cascade resilience is counter proportional to its spectral radius and vulnerability (probability of a node failing because its neighboring node has failed). As spectral radius and vulnerability increases, cascade resilience decreases. A cascade resilience measure is proposed that ranks a network with a certain spectral radius and vulnerability on a scale of 0 to 10.
- *Fundamental resilience*: In general, and assuming random faults, cascade resilience versus vulnerability defines a straight line called the fundamental resilience line. The network's *critical vulnerability* is equal to the point along the horizontal vulnerability axis where the line crosses zero on the vertical axis. This has profound implications: cascade failures within complex CIKR systems increase in severity and frequency with increases in spectral radius and vulnerability. Networks whose resilience line is below zero are subject to catastrophic collapse.
- *Targeted attacks*: Risk and resilience of a complex CIKR network vary according to which node/link is targeted for an attack or fault. Critical nodes with high influence and connectivity act as *super-spreaders* that magnify consequences. Critical links with high betweenness (or blocking criticality) may reduce resilience because they are more likely to break when the network is under stress. Targeted attacks on critical nodes can transform a network from normal accident prone to complex catastrophe prone.
- *Network flow*: Similar results are obtained for networks that deliver a commodity such as water, gas, oil, electrons, or packets through a flow network. Criticality factors based on betweenness are appropriate measures of bottlenecks in flow networks. Consequence in these networks is defined as the *loss of flow* due to a single node or link failure. Unfortunately, Braess's paradox may lead to contradictory results because flow may actually increase when a node or link fails. In this case, network flow exceedence may be a misleading measure of resilience. Multipath betweenness, blocking links and nodes, and an analysis of rerouting through alternative paths may be better measures for flow networks.
- *Robustness*: A robust network is one that can withstand damage to nodes and links and remain in one piece. Link robustness is a measure of how many links can be removed before separating a connected network into disconnected components or islands. Node robustness is a measure of how many nodes can be removed to separate a connected network into components. Robustness may actually work against cascade resilience by increasing SOC. If robustness is favored over cascade resilience, link and node robustness is a better metric of resilience.
- *Network risk*: The definition of PRA risk is extended to networks by summing weighted TVC values across all nodes and links. Weights may be the product of one or more criticality factors: vulnerability, connectivity, influence, betweenness, overloading, and blocking. For example, a weight might be obtained by multiplying network properties such as connectivity and betweenness together. Weights are used to shift resource allocation to the most critical nodes and links, especially when resources are limited.
- *Fragility Framework*: Fragility is the opposite of resilience. A fragile system is more likely to suffer more frequent and more consequential events than a resilient system. In terms of whole-of-community resilience, the Hodges Fragility Framework provides a more comprehensive model of community resilience across a broad scope of people, organizations, systems, and budgetary constraints. It rightfully embraces complexity theory as described in Chapter 3, and it can be extended to semi-quantifiable measures of resilience. It is a broader way to characterize resilience—or lack of it—than network science, but it is also much more difficult to quantify and optimize.

4.1 CIKR AS NETWORKS

An interrelated, interdependent, and linked CIKR system is represented as a *network* containing *nodes*, *links*, and a *topological map* or wiring diagram. Figure 4.1 illustrates two different systems represented as complex networks. Nodes and links can be anything desired in the model. The topological map defines how nodes are connected to one another and defines the topological structure of the system. In Figure 4.1a nodes and links are physical assets of drinking water system. In Figure 4.1b, nodes are human actors and links are their social network ties.

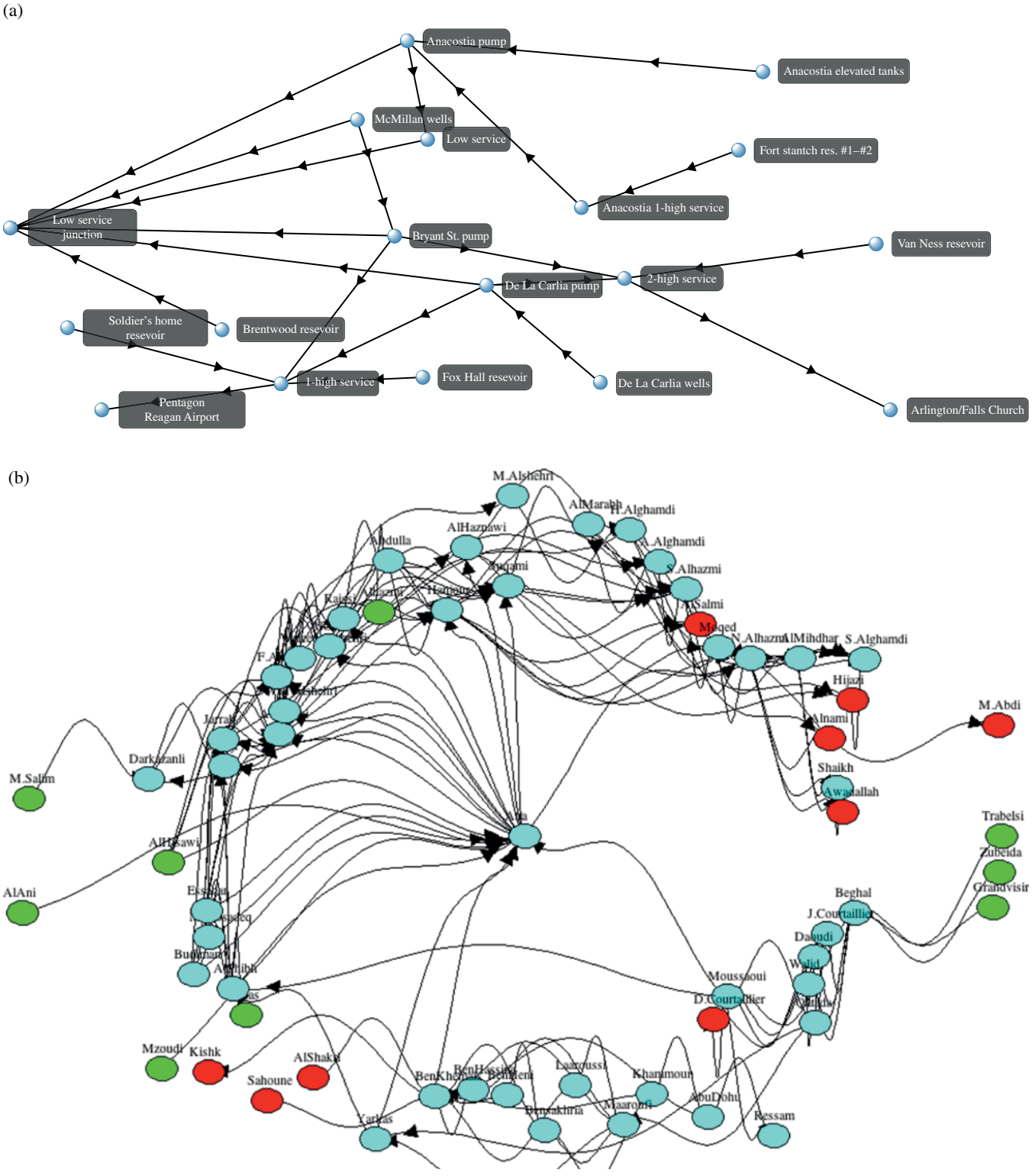


FIGURE 4.1 The complex CIKR system network of the Washington, DC, drinking water system illustrates the application of network science to a physical system, while the system network of the al Qaeda 9/11 terrorist organization illustrates the application of network science to relationships. (a) Washington, DC, water network containing pipes, intersections, reservoirs, and pumping stations. (b) 9/11 terrorist organization network containing individuals and their social connection to one another.

Networks are not unlike road maps connecting cities together. Each city is a node and the roadways are links. A road map shows the connections and therefore is a topological map of the road network. But a roadmap can also be an energy pipeline, communication, power, waterway, or rail map. *Network science* is simply a technology for modeling a complex system as nodes, links, and connections. It is perhaps the *simplest* technology for representing complex CIKR systems.

The utility of network science is that nodes can be anything—Internet switches, bridges, hospitals, police stations, airports, cell phones, and so on. Links can be physical (roadways, pipes, wires, or power lines) or relational (members of the same social club, links to Web pages, or influences of policy). Figure 4.1 illustrates both types of networks—the physical water system of Washington, DC, and the relational 9/11 terrorist network.

The topological structure of a network can be thought of as a list of connections such as railway station A connects to station B through link number 5 or person A knows person B. In the formal connection matrix definition, only one link is allowed to connect pairs of nodes. As it turns out, topology plays an important role in resilience, because some topologies are more resilient than others. A network's spectral radius is an overall measure of SOC or topology in general. Higher values of spectral radius mean higher levels of SOC, which in turn means larger cascade collapses. Details can be found in Appendix C.

Figure 4.1a contains major components in a drinking water network consisting of reservoirs, pumping stations, and loads (customers). Pipes and river flows are represented as links. Links are directional in Figure 4.1a, but they can be *bidirectional*, meaning they carry a commodity such as water in both directions. A *directional* network contains one-way links. For example, the De La Carlia wells feed into the De La Carlia pumping station, so the directional topology of these two nodes is

$$\text{Del La Carlia Wells} \rightarrow \text{De La Carlia Pump}$$

Figure 4.1b is a *social network* containing known members of the al Qaeda cell that planned and carried out the infamous 9/11 terrorist attacks on the United States. The links represent social connections—who knew whom—before the attacks took place. At the center of this social network is the mastermind leader Mohammed Atta. Atta is considered the leader because he is the hub of this network: he has more connections—22—than any of the others. The *degree* of a node is equal to the number of connections, and the *hub* of a network is the node with the largest *degree*. Degree is a measure of connectivity. To obtain a node's connectivity, simply count the number of links connecting it to the rest of the network.

Atta is also the maximum *betweener*, because his *betweenness centrality measure* is higher than all other nodes and links. Betweenness is a laborious calculation best done by a computer. It is defined as the number of paths running through a node or link *from* all other nodes *to* all other nodes of the network. Betweenness is related to *bottlenecks* because high betweenness centrality often indicates a potential bottleneck when there is a surge in the flow of a commodity through the network. Bottleneck and betweenness are interchangeable terms for potential congestion in a flow network.

Betweenness is a criticality factor because of its causal relationship with bottlenecks. A high betweenness value of a node or link means the node/link is more critical than other nodes/links, because its removal or blockage would reduce flow through the network more than lower-valued nodes/links. The De La Carlia pump node is the highest ranked betweenness node in Figure 4.1a, while the Low Service Junction is next in rank. Actor Atta has the highest betweenness in Figure 4.1b with 2519 paths running through this central node. Moussaoui and Ben Khemais rank second in terms of betweenness, while N. Alhazmi ranks third.

A node or link is considered *blocking* if its removal segments the network into isolated islands. Blocking is an extreme form of betweenness or bottleneck because removal of a blocking node/link can potentially block the flow of a commodity from one isolated component to another. The drinking water system in Figure 4.1a contains six blocking nodes: Anacostia pump, Anacostia 1-high service, Low Service Junction, De La Carlia pump, 1-high service, and 2-high service. Ten of the 18 links are blocking links. Figure 4.1b contains five blocking nodes (Atta, Ben Khemais, Darkazanli, Begal, and N. Alhazmi) and five blocking links—all connected to one of the blocking nodes.

A measure of influence called the eigenvalue determines a node's impact on all other nodes in terms of cascading. Obviously, a highly connected node has an enormous influence on its neighboring adjacent nodes. But influence is even greater if the neighboring nodes are also highly connected. The influence of a node on all others is transmitted via the connectivity of its neighbors, the neighbor's neighbors, and so on. Influence is simply the eigenvalue of a node, which can be calculated as described in Appendix C. Spectral radius is the largest eigenvalue across all nodes. In Figure 4.1a, Low Service Junction is both the hub with the most connections and the most influential node. De La Carlia pump and Bryant St. pump are the second most influential nodes. In Figure 4.1b, Atta is both hub and most influential, but Al Shehhi is the second most important node.

There are other network properties of interest to CIKR systems besides connectivity, betweenness, blocking, and influence. For example, the *diameter* of a network is the maximum number of hops—steps along a chain of links from one node to another—needed to travel from any node

to any other node. Diameter is a *small world* metric, because the diameter of a small world network grows more slowly than its number of nodes.[2] The diameter of the 18-node network in Figure 4.1a is 6 hops. The diameter of the 62-node al Qaeda network in Figure 4.1b is 5 hops. Therefore, the al Qaeda network is smaller than the DC water network even though it has three times as many nodes. Why?

In general, as the number of connections per node increases, the diameter decreases. The DC water network has 22 links for a ratio of 2.44 links/node, while the al Qaeda network has 150 links for a ratio of 4.83 links/node.[3] The al Qaeda network is denser and therefore has a slightly smaller diameter. This may seem like an insignificant detail, but network density is one of the major factors in network collapse and is related to *percolation*—the process of adding links to a network. Percolation increases the *SOC* of a network, which makes some networks more fragile.

Another property of a network is called its *cluster coefficient*. This is a measure of how many neighbors of nodes are connected to one another. For example, a family consisting of mother, father, sons, and daughters forms a cluster because mother, father, and siblings are related by familial links. In contrast, unrelated mothers, fathers, and siblings are not linked by familial connections. The cluster coefficient of the related nodes is 100%, and the cluster coefficient of the strangers is zero. Clustering plays a big role in social network where it directly relates to the spread of memes. Memes are neither good nor bad, unless a "fake news" meme is misunderstood as truth and misinformation influences presidential elections and political outcomes.

4.1.1 Emergence

Complex CIKR networks are not static structures. Instead, they evolve over time through a process of adaptation called *emergence*. Emergence is a dynamic process of adding and rewiring nodes and links to respond to outside forces and increase system fitness.[4] More formally, mapping is a function of time, which changes as links and nodes adapt to environmental factors—threat, efficiency factors, regulation, cost, NIMBY, and demand. *Preferential attachment* is one example of an adaptive and emergent process. It favors one node over all others in response to economic, efficiency, or other forces.

Preferential attachment—as illustrated in Figure 4.2—is the most common form of emergence. Here is how it works: as nodes and links are added to a growing network such as an expanding Internet, drinking water system, highway network, or power grid, *competitive exclusion* begins to favor one node over all others. Figuratively, links prefer this node. As the favored node gains links, it becomes even more attractive to other links simply because it has the most links. More links beget even more links, which results in one dominant node with far more links than all others.

Figure 4.2a shows a random network created by randomly connecting pairs of nodes. Figure 4.2b shows a scale-free network created by preferential attachment. Note the difference in structure. The random network has no apparent structure, but it actually has a *random structure*, because the distribution of links to nodes obeys a *binomial distribution*, which is a consequence of random selection. This distribution is identical to the distribution obtained by tossing a coin many times and counting the number of heads. On average, the number of heads will be equal to the number of tails. The probability of k heads in n tosses obeys a bell-shaped binomial distribution. Accordingly, a random network is formed in much the same way: k links are attached to each node in n trials. Therefore, the number of links attached to each node obeys a binomial distribution too. Figure 4.2a shows one possible random network produced by making random connections and its corresponding link distribution. Note that it is symmetrical around mean degree of 4 links/node. (200 links equals 400 connections, because a link connects a pair of nodes. Therefore, 400/100 equals 4.)

The *scale-free network* of Figure 4.2b is a different story, because node selection is biased according to preferential attachment. Think of this biased process as tossing an unfair coin that is more likely to turn up heads than tails as the number of heads already turned up in the past changes the odds. The first toss produces either a head or tail, but the second toss produces a head with higher likelihood if the first toss produced a head. The third toss produces a head with probability of 50% if the previous 2 tosses produced 1 head and with probability of 67% if the previous 2 tosses produced 2 heads. As the number of links attached to a node increases, the likelihood of additional links attaching to the same node increases.

The distribution of links to nodes obeys a *power law* because of this bias. The distribution shown in Figure 4.2b is a long-tailed power law with fractal dimension of 1.88. The scale-free network derives its name from the fact that its link distribution is scale-free, for example, is a self-similar fractal. Like fractals in general, power laws remain shaped like a power law at all scales. However, this distribution should not be confused with the fractal distribution of hazards, as described in Chapter 3 even though they are both power laws.

The fractal dimension of the power law produced by preferential attachment emergence depends on the density of the network. As density increases, so does the fractal dimension. Furthermore, the diameter of a scale-free network rapidly *declines* as density increases (along with fractal dimension). Therefore, a scale-free network is also a small world. In

[2]Diameter grows proportional to the logarithm of number of nodes.

[3]The mean degree of a network is 2(#links)/#nodes.

[4]Fitness is the ability of a system to maintain or increase its survivability now and in the future.

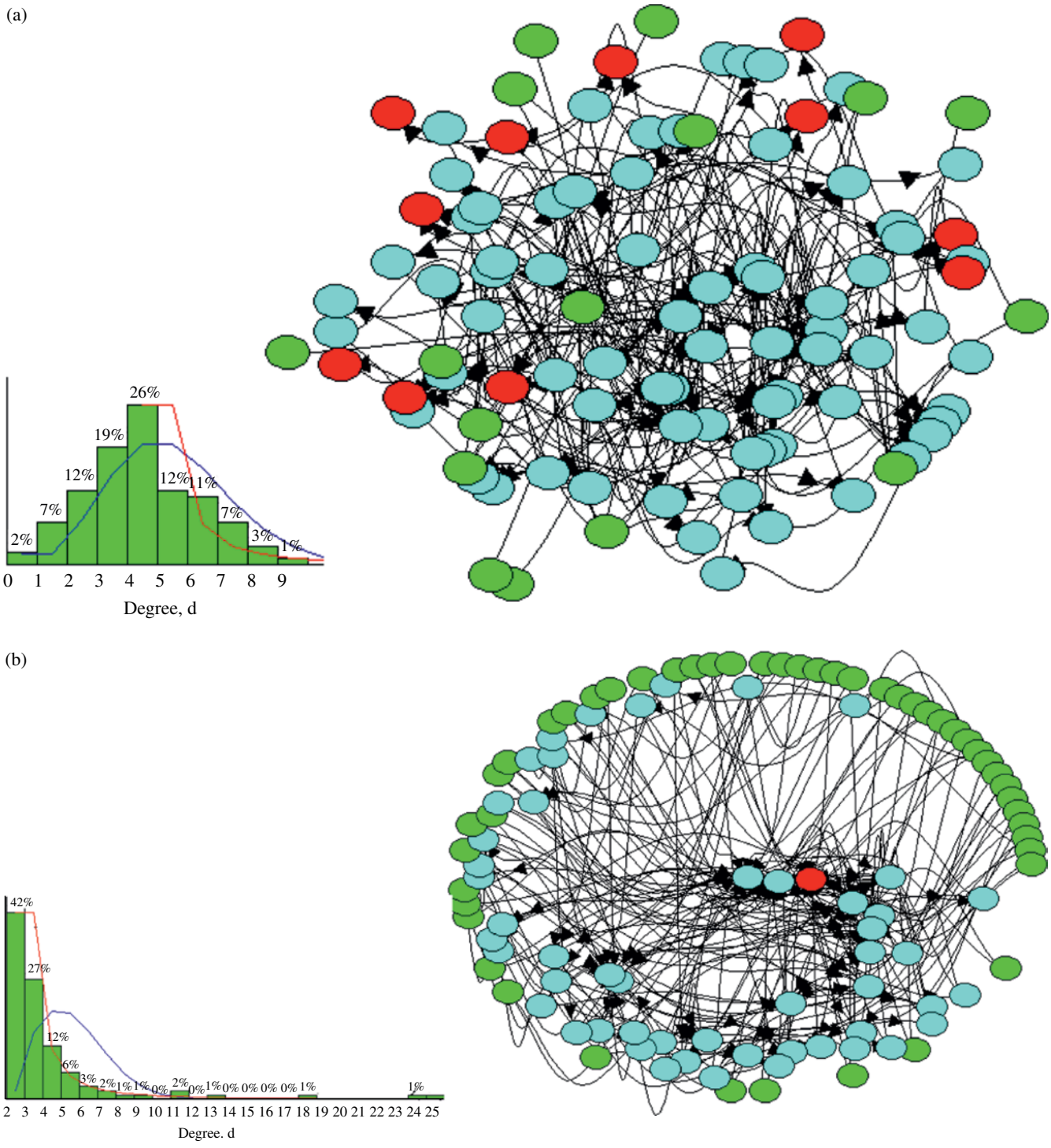


FIGURE 4.2 CIKR networks are typically random, scale-free, or clustered. (a) Network of 100 nodes and 200 randomly connected links shows not apparent structure. The distribution of links to nodes obeys a binomial distribution. (b) Network of 200 links and 100 preferentially selected nodes—a few nodes have many connections, but most nodes have a small number of connections. The distribution of links to nodes obeys a long-tailed power law. (c) Network with 9 nodes and 12 links. Two clusters containing nodes with high cluster coefficients are connected by a betweener node and two betweener links, each with 60 paths. A cluster coefficient greater than 1.0 indicates high connectivity of adjacent nodes.

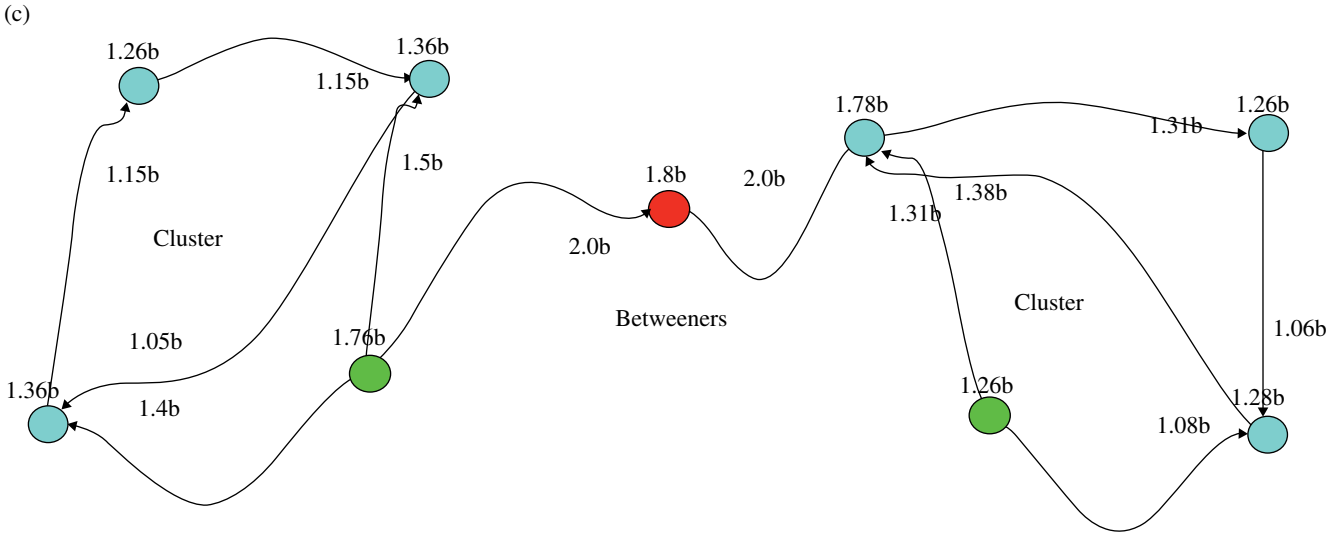


FIGURE 4.2 (Continued)

Figure 4.2b, fractal dimension increases from 1.88 to 1.99 as mean degree of the network increases from 3.9 to 15.3 links/node. Diameter declines from 6 to 3 hops. That is, one-half as many hops are needed to reach any node from any other node.

The diameter of a network is important because normal accidents propagate faster and further in a small world than a large world network. Faults can spread to all parts of the system with fewer hops in a smaller world. But as shown, the diameter is determined by link density. As more links are added to a fixed number of nodes, link density increases, which reduces the diameter. The process of adding links is called *link percolation* and is one of the sources of SOC.

As it turns out, density, as measured by mean degree, is not a particularly useful measure of network complexity, because it ignores the underlying structure of a CIKR network. For example, the mean degrees (density) of the random and scale-free networks of Figure 4.2 are identical, and yet, these are clearly different structures. Density is inadequate to categorize a CIKR network. Therefore, a different metric is used. *Spectral radius* is a measure of both density and structure.[5]

The spectral radius of the scale-free network in Figure 4.2b is 7.9 compared with 4.7 for the random network of Figure 4.2a. The mean degree of both is 4.0, so why the difference in spectral radius? The scale-free network is more structured, because it has a large hub with high degree and many smaller-degreed nodes. This structure shows up in the spectral radius. Spectral radius is a better measure of SOC than density, because it measures density *and* structure.

The network in Figure 4.2c is much different than the previous examples, because it has clusters and high-betweenness nodes and links. Its spectral radius is also much different than the random and scale-free networks. Actually, its spectral radius is approximately equal to its mean degree—2.51 versus 2.66. Instead of degree structure, this network exhibits betweenness and cluster structure. The cluster coefficient of a node is equal to the ratio of number of connected neighbors to node degree. In Figure 4.2c, cluster coefficient is normalized so that a coefficient of 2.0 means maximum clustering, while a coefficient of 1.0 means zero clustering.

Link betweenness is a rough measure of the criticality of links, because their removal eliminates more paths through the network than removal of a lower-valued link. Therefore, betweenness is especially valuable when considering the loss of network flow. That is, pipeline, power line, and communication systems that carry water, gas, oil, electrons, and packets depend on links with high betweenness more than lower-valued links. High-betweenness nodes and links have been identified as critical in the US Western power grid. High-betweenness links lie on critical paths in the grid and lead to its separation into isolated components when they fail. The Western power grid collapsed for 4 h in 1996 due to the failure on a single power line near the Oregon–California border. This complex CIKR is studied in detail in Chapter 13.

4.1.2 Classes of CIKR Networks

Figure 4.2 illustrates the three main classes of CIKR networks studied in this book. Random networks are rarely found in practice, but they are more resilient against single node/link attacks or failures, because all nodes and links are approximately the same. Nodes and links have similar degree and betweenness, and the spectral radius of a random network is low. Random networks are the most resilient class of network.

[5]Spectral radius is the largest nontrivial eigenvalue of the connection matrix.

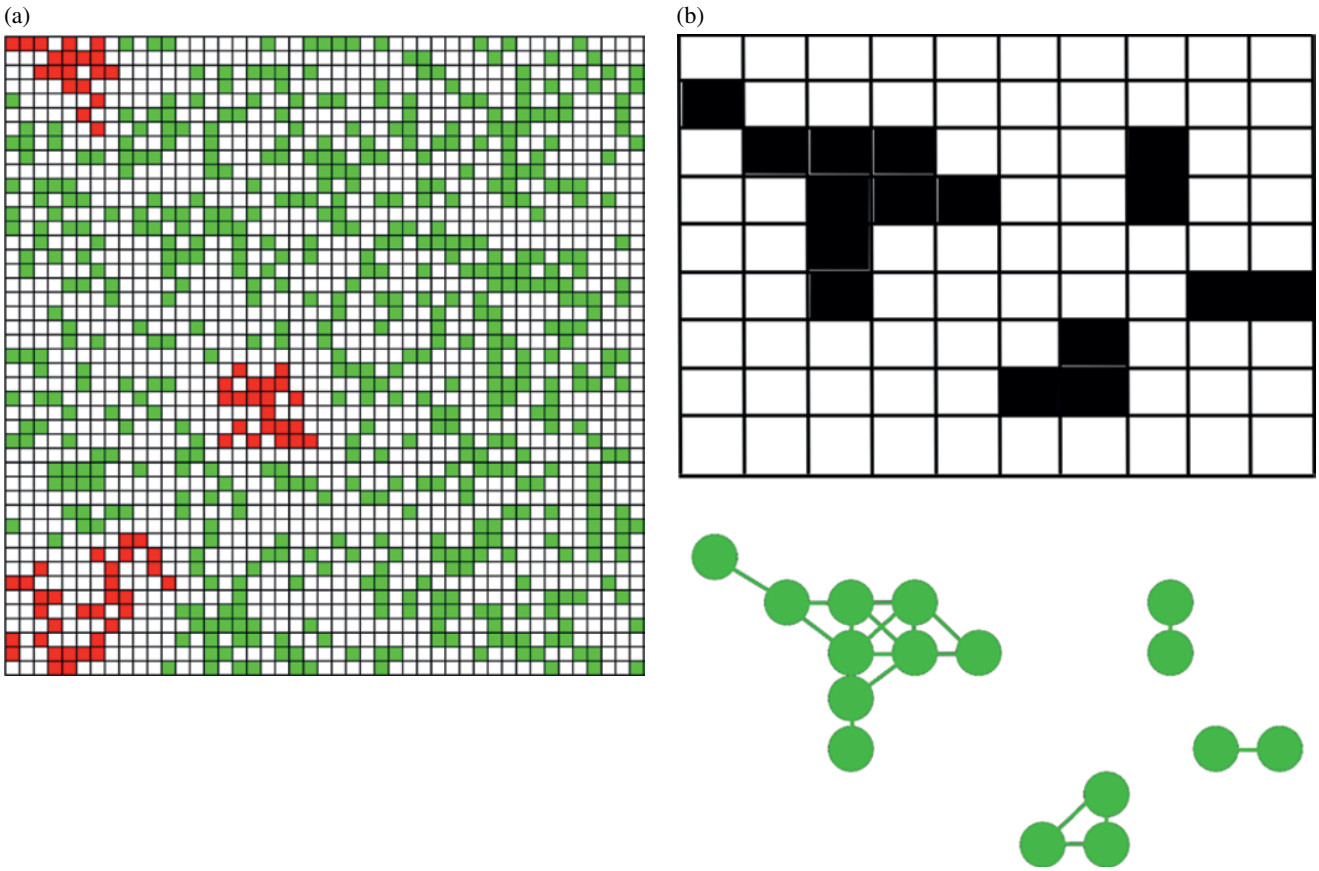


FIGURE 4.3 Forest fires are simulated as a grid containing cells that are either empty (white space) or contain a tree (shaded space). The grid is converted into a network by linking adjacent tree-containing cells together. (a) A forest grid containing empty cells or cells with trees. When a tree catches on fire, it is dark. (b) The rule for translating the forest grid into an equivalent network is to draw links between adjacent cells containing trees. Nodes correspond with trees and links correspond with adjacency.

Scale-free networks are the opposite: their structure makes them more resilient in general and less resilient in particular. In general, they are less vulnerable to random attacks, because most nodes and links in a scale-free network have few links and low betweenness values. On the contrary, they are specifically more vulnerable to targeted attacks against their hubs and betweeners for obvious reasons. Spectral radius is higher for equally dense scale-free networks; so SOC exists in the form of hubs and betweeners. Therefore, hubs and betweener nodes and links are the Achilles' heel of networks and should be protected against targeted attacks.

Clustered networks are somewhere in between random and scale-free networks in terms of fragility. In general, clustered networks are nearly as resilient as random networks, because a clustered network's spectral radius is low. However, clustered networks may be easily separated into isolated clusters by destroying a blocking node, critical link, or high-betweenness node or link. Figure 4.2c illustrates this vulnerability. Removal of one or two betweener links or nodes divide this network into isolated clusters called islands. If this were a water supply network, separation means water cannot reach parts of the network. In Figure 4.1a, removal of the top two betweener nodes (De La Carlia pump and Low Service Junction) separates the Washington, DC, water network into four isolated islands.

4.1.3 Self-Organized Networks

Thus far percolation, connectivity, betweenness, and influence have been identified as factors that increase SOC in networked infrastructure systems. They are measures of causality, because high values of these factors have consequences in terms of cascading or loss of commodity flow when the network is stressed. They may seem abstract and meaningless when applied to various CIKR systems and corresponding hazards, but they are the keys to understanding fragility and risk. The classical *forest fire* simulation shown in Figure 4.3 demonstrates how percolation, spectral radius, and self-organization all fit together and contribute to fragility and risk.

Consider a two-dimensional grid representing a forest, as shown in Figure 4.3a. Each cell of the grid represents a plot of land that either is empty or contains a tree. The forest is threatened by occasional lightning strikes on either an empty cell or one containing a tree. When a tree is struck by lightning, it bursts into flames and burns to the ground, leaving the cell empty once again. In addition, adjacent trees, if they exist, are ignited and also burn to the ground. Forest fires spread through igniting adjacent trees until reaching an empty cell that stops further conflagration.

This simple simulation is another metaphor for a complex system subject to SOC. It is simple, and yet it behaves in unexpected ways. Suppose, for example, the simulation is repeated thousands of times: trees are randomly planted in empty cells, and random lightning bolts strike at regular intervals and either do no harm or ignite a tree. The ignited tree spreads flames to adjacent cells containing trees that also burn down. The conflagration from adjacent cells continues until there are no more adjacent cells containing a tree. This scenario is repeated thousands of times and the exceedence probability (EP) distribution constructed. What is its shape and what does it say about risk and resilience?

Forest fires have consequences beyond the number of trees destroyed, but for simplicity, assume consequence is equal to the number of trees destroyed by the conflagration following a successful lightning strike. We want to determine the PML risk associated with lightning strikes. Recall that PML risk is an entire distribution, not simply a single number, although it also produces maximum PML, which is a single number. The shape of this distribution is captured in a single parameter—the fractal dimension produced by the EP distribution of number of trees catching fire. Similarly, risk is captured in a single number if we use maximum PML as a measure of system risk due to cascading collapse of the forest.

Recall that the PML risk curve is obtained by multiplying together vertical and horizontal axes values of the exceedence distribution. This is shown as a dark graph rising above the EP graphs of Figure 4.4. Maximum PML is the value of the PML graph at its maximum point. In Figure 4.4, maximum PML is 8.25 for frequent lightning strikes and 9.88 for less frequent strikes. That is, risk is lower for the scenario where lightning strikes more often!

Fractal dimension is related to resiliency of the forest. Recall that resilience increases with an increase in fractal dimension. Figure 4.4 shows simulation results for two scenarios. In the first scenario, a bolt of lightning strikes the forest once every 20 months, yielding a fractal dimension of 0.58. In the second scenario, the lightning strikes once every 40 months, yielding a fractal dimension of 0.48. Accordingly, the frequent lighting strike scenario is less resilient than the infrequent or less frequent scenario.

Scenario two—infrequent strikes—is higher risk than scenario one, even though the lightning bolts strike half as often. How can fewer strikes *increase* risk? Note that the power law approximation of infrequent strike EP has a heavier or fatter tail than the power law for the frequent strike EP curve. Less strikes means less resilience! How so? A longer interval between strikes allows the forest to become more self-organized, where the definition of forest self-organization is dense growth of trees. Longer intervals between lightning strikes means more time for the forest to self-organize. A highly organized forest is a dense forest. In terms of network science, adding trees is equivalent to percolation, which is a form of self-organization.

Figure 4.3b shows how to transform the forest fire grid into a network. Each tree is represented as a node, and adjacency is represented by a link connecting adjacent trees. Empty cells in the grid are ignored, and nonadjacent trees have no connecting link. Thus, the forest fire simulation is identical to a network, and the fire is identical to a cascade failure. For this reason, both grid and network models are contagion models of cascade failures. The collapse of a power grid, destruction of a forest by a fire, spread of a deadly virus among humans, or spread of malware throughout the Internet are all contagion models of cascade failures.

The forest fire simulation illustrates how percolation increases SOC. Longer intervals between lightning strikes allows more time for trees to grow (percolation), which increases the size of tree clusters (self-organization) so that when a lightning bolt strikes, it has more fuel to burn. By increasing the frequency of strikes, the size of each burn is reduced. Similarly, if a network model of the forest system is used, long intervals of time allow for more nodes and links to join the network (self-organization), which increases the spectral radius of the network. And we already know from the Chapter 3 that spectral radius is a measure of fragility.

Percolation of nodes and links increases risk and reduces resiliency because it increases the spectral radius of a connected system. As nodes and links join the network (forest), they increase SOC, which decreases resilience and leads to greater PML risk.

Percolation and SOC:
Percolation increases risk by increasing the spectral radius of a CIKR system represented as a network.

4.2 CASCADING CIKR SYSTEMS

Network cascades are like landslides in Bak's sand pile model of complex CIKR systems. They can be even more disastrous than single-asset failures, because of dependencies across sectors. For example, a power grid outage can spread to natural gas and oil supply networks that depend on electricity to run pumps and process gas and oil products. The outage can also impact water systems

that depend on electric power to run treatment facilities. The horrific Fukushima Daiichi power plant meltdown was exacerbated by the tsunami wave because it destroyed the power lines into the power plant. Without electric power from outside of the power plant, cooling failed, and without cooling, the reactor core overheated. The Fukushima disaster was a classical normal accident with Bak sand pile behaviors. Accordingly, its EP distribution curve would have been long-tailed just like the graphs in this chapter.

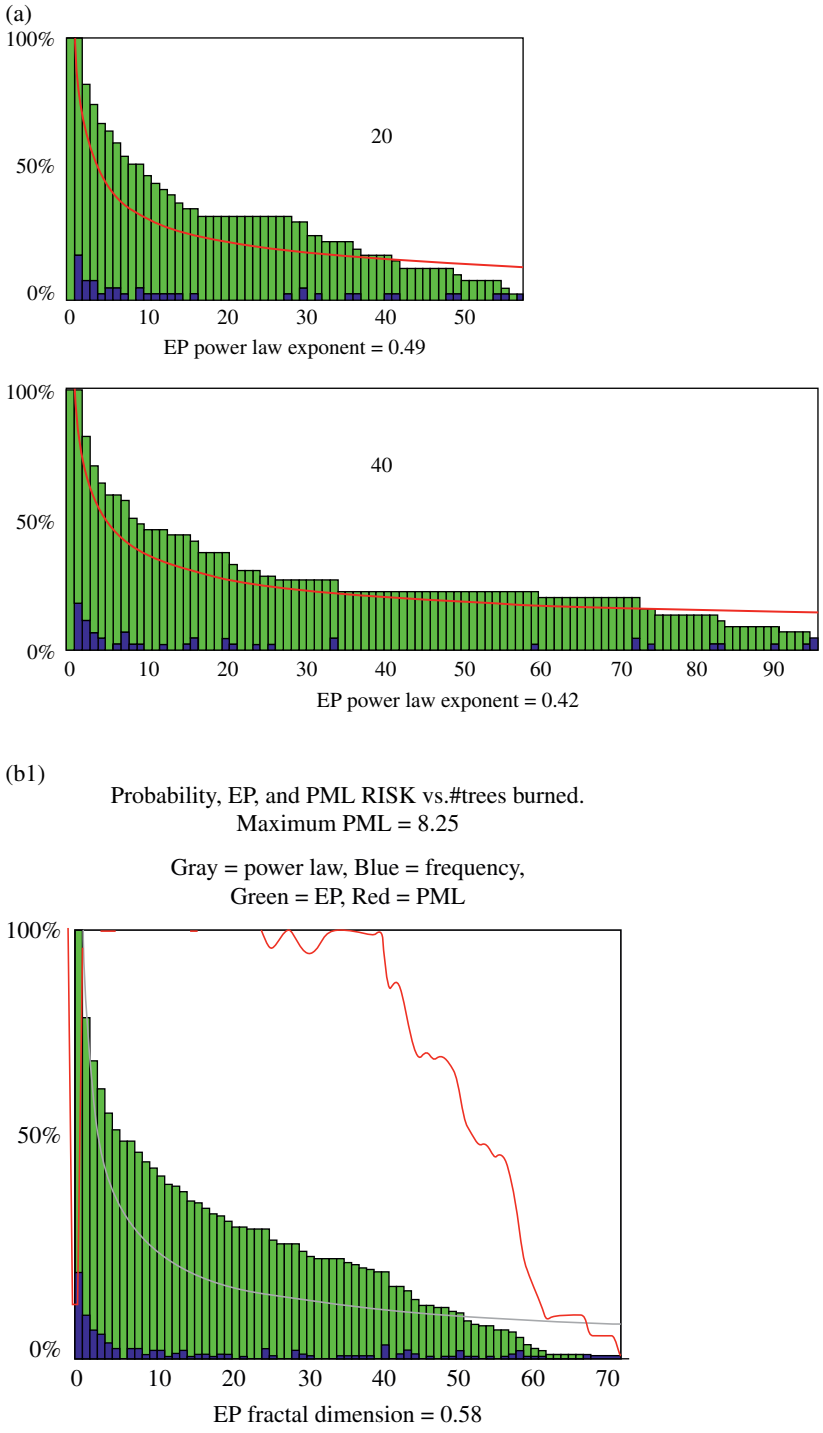


FIGURE 4.4 Results of the forest fire simulation for two scenarios: (a) frequent lightning strikes once every 20 months, and (b) infrequent/less frequent lightning strikes once every 40 months. For 20 months, fractal dimension of the exceedence probability equals 0.58, and PML risk equals 8.25 trees; and for the 40-month scenario, fractal dimension equals 0.48, and PML risk equals 9.88 trees per strike.

Network cascades behave like collapsing sand piles, and their risk and resilience profiles are typically long-tailed.

As an example of the spread of a fault through a networked CIKR, consider the power grid network in Figure 4.5. Like dominos falling in a chain reaction, a cascade failure starts with a single node failure and propagates to adjacent nodes with a given probability called vulnerability here. Each connected neighbor fails with probability equal to its vulnerability as the fault spreads to adjacent nodes, and so forth. For obvious reasons, this is known as the contagion model of cascade failure.

Figure 4.5 shows the result of simulated cascades. An initial node called the *seed* is tripped, which causes connected nodes to also fail with probability equal to the node's vulnerability. This chain reaction continues until all nodes have failed or the chain reaction dies out on its own. Obviously, higher vulnerability means the chain reaction propagates further. Self-organization as determined by a network's spectral radius also contributes to the spread of failures. Higher spectral radius means the chain reaction impacts more nodes. Therefore, cascade failures and resilience against failures depend on two factors: the vulnerability of each node and the spectral radius of the network.

Figure 4.6 compares two exceedence probabilities for the same network shown in Figure 4.5 but for different vulnerabilities. The results for vulnerability of 25% fits a power law with fractal dimension of 2.45 and yields a maximum PML risk of 0.47 failed nodes. The results for vulnerability of 50% are dramatically different. The power law fit yields a fractal dimension of 0.44 and a maximum PML of 6.5 nodes. This clearly illustrates the relationships among node vulnerability to cascading, fractal dimension, and PML risk.

(b2)

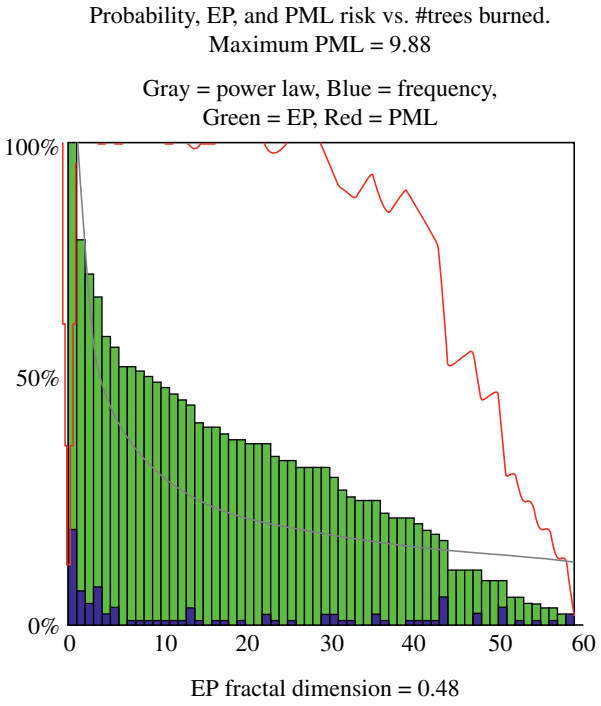


FIGURE 4.4 (Continued)

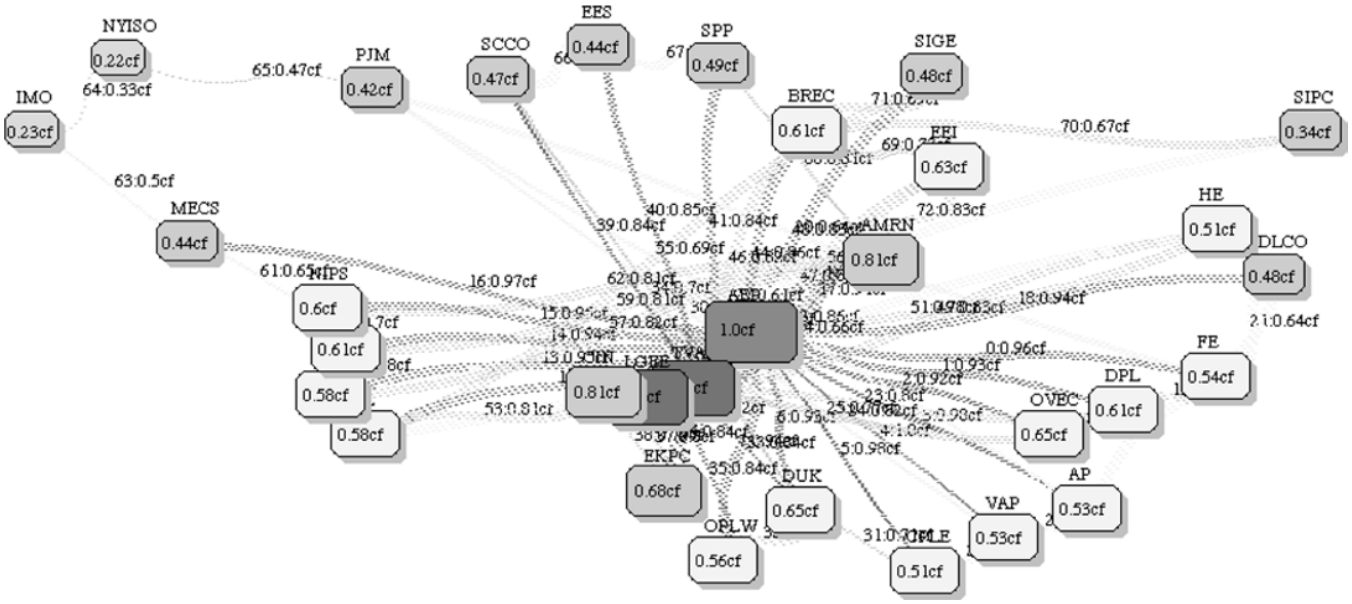


FIGURE 4.5 This cascade frequency heat map of a small microgrid section of the power grid shows the frequency of failures of nodes due to a cascading failure originating from a single randomly selected node. Most likely nodes to fail due to cascading gravitate toward the center and are darker color, while less likely to cascade nodes are pushed outward and lighter colors. The number in each node (and links) is the frequency of failure.

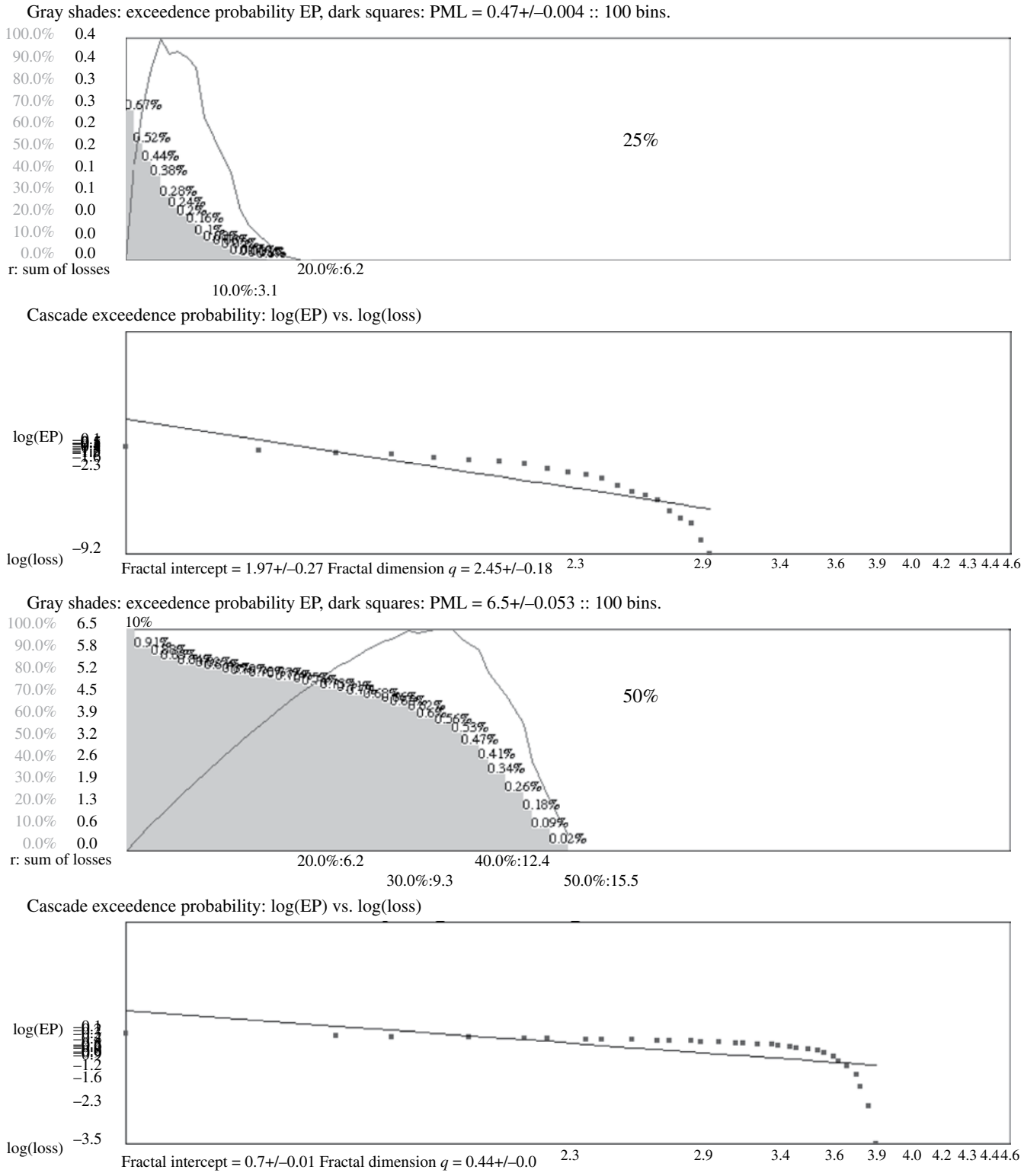


FIGURE 4.6 Exceedence probability for small vulnerability network nodes obeys a power law, but for large vulnerability, the power law applies only to low-consequence events. High-consequence events obey a normal distribution.

TABLE 4.1 Cascade failure results for increases in parameters spectral radius, vulnerability, and fractal dimension

Result	Spectral radius	Vulnerability	Fractal dimension
Max. PML risk	Up	Up	Down
Resilience	Down	Down	Up
Fractal dimension	Down	Down	

Similar results can be obtained by comparing two networks with different spectral radii. Recall that spectral radius is a measure of self-organization, and in general, a higher spectral radius means higher cascade risk and lower resilience. In fact, these relationships are summarized in Table 4.1, where an increase in parameters produce an up/down change in the results listed in the first column. For example, an increase in vulnerability produces an increase in cascade PML risk and decrease in resilience and fractal dimension.

Figure 4.6 illustrates an important difference between severities of cascading failures. Note that the exceedence distribution of the simulation results with vulnerability of 25% is relatively well behaved. It is a near perfect power law with fractal dimension much greater than 1.0—2.45 actually. The tail of the power law is short and the maximum PML risk is low. This is a simple cascade failure with mathematically regular properties.

Note that the exceedence distribution for the simulation results with vulnerability 50% is not well behaved. In fact, the EP curve is not a power law because its tail drops off too slowly. This is reflected in the fractal dimension, which is much less than 1.0—0.44 actually. Furthermore, its maximum PML risk is much greater at 6.5 collapsed nodes—an order of magnitude larger than the well-behaved simulation.

Figure 4.6 illustrates a phase transition from well-behaved collapse to catastrophic collapse. The network crossed a *critical point* defined by fractal dimension of 1.0. When fractal dimension exceeds 1.0, the exceedence distribution shifts from a pure power law to a curve that is more normal than a power law. Clearly, risk has also crossed a critical point from mild to catastrophic.

Complex Catastrophe Critical Point:
A highly connected, interdependent complex CIKR system transitions from low risk to high risk when the fractal dimension of its exceedence probability curve exceeds 1.0 and transitions from normal accident to complex catastrophe when the product of vulnerability and spectral radius is much larger than 1.0.

4.2.1 The Fundamental Resilience Line

Figure 4.7 shows the relationship between fractal dimension and the product of vulnerability and spectral radius for a number of complex CIKR systems studied in this book. The vertical axis is the logarithm of fractal dimension. The horizontal axis is the product of vulnerability and spectral radius. Note that the relationship is linear. That is, the points for a given network fall on a straight line. This is the fundamental resilience line defining resilience of cascading networks.

The best-fitting straight line for these networks of varying sizes and sectors yields a *fundamental resilience equation* as described in more detail in Appendix C. Basically, the equation establishes a linear relationship between fractal dimension and the product of vulnerability and spectral radius called cascade resilience. The zero crossing point defines a critical vulnerability.

Critical Vulnerability:
A network's critical vulnerability point is the value of vulnerability to cascading where the fundamental resilience line crosses zero.

Table 4.2 lists the critical vulnerability points for the 11 networks analyzed in Figure 4.7. The average critical vulnerability is 14.9%. This point separates a resilient network from a fragile network.

A normalized measure of resilience can be calculated by noting that the lines in Figure 4.7 cross have different slopes and critical vulnerability points. A steep slope suggests that a network is less resilient than a network with a shallow slope. But the vertical intercept also impacts resilience. Both must be considered, including the critical vulnerability value. Figure 4.8 illustrates the role of these parameters in determining cascade resilience versus node vulnerability for a given network with a given spectral radius. Resilience is normalized to an interval between 0 and 10, as suggested by Figure 4.8.

The objective of CIP is to increase resilience and reduce risk. This means fractal dimension should be as large as possible, and spectral radius should be as small as possible. Restructuring or rewiring a network's connections may reduce spectral radius. While restructuring is beyond the scope of this book, note the following:

Cascade resilience depends on network topology as measured by spectral radius and asset vulnerability as measured by the probability an adjacent node/link will fail. Resilience increases with a decrease in both—spectral radius and vulnerability.

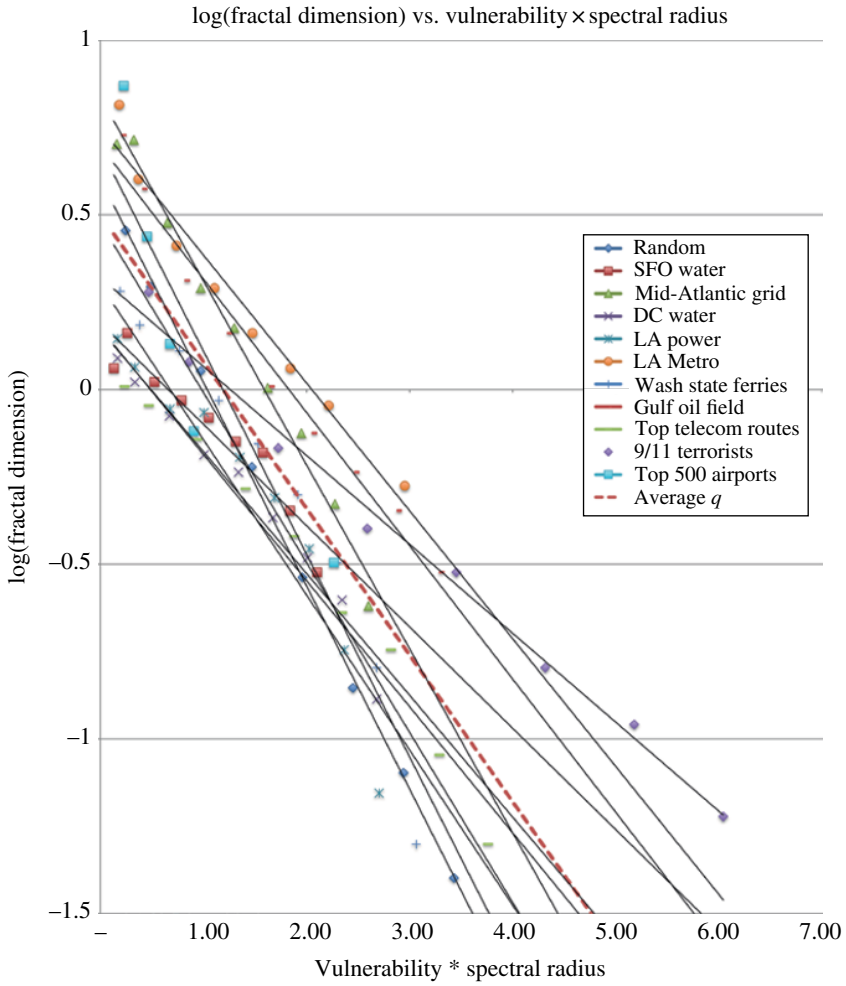


FIGURE 4.7 Fractal dimension of cascade failures in a CIKR network declines exponentially with respect to the product of asset vulnerability and spectral radius.

TABLE 4.2 Constants for the 11 networks analyzed in Figure 4.7 and critical point vulnerability indicates that the top 500 Airport network is the most fragile

Network	Spectral radius, ρ	b	k	Critical point, γ_0 (%)
911 terrorist	8.63	0.32	–0.25	14.8
Mid-Atlantic power grid	3.25	0.84	–0.53	48.8
LA Metro	3.69	0.75	–0.37	54.9
Random	4.90	0.61	–0.58	21.5
Top 500 airports	45.30	0.69	–0.58	2.6
DC water	3.35	0.17	–0.35	14.5
LA power grid	3.38	0.30	–0.44	20.2
Top 30 telecom routes	4.70	0.18	–0.36	10.6
Wash State Ferries	3.83	0.48	–0.48	26.1
SFO water and power	2.63	0.19	–0.29	24.9
Gulf oil field	4.11	0.70	–0.38	44.8
Average	7.98	0.50	–0.42	14.9

4.2.2 Critical Factors and Cascades

All nodes and links are not created equally. Some are more critical than others, because they contribute more to cascading than others. What factors matter? To find out, the following simulations were performed as before while correlating cascade frequency with three criticality factors: connectivity, influence, betweenness, and blocking node status.

Connectivity is equal to the number of links a node has; influence is a measure of the influence a node has on its neighbors, the neighbor's neighbors, and so on; betweenness is normalized number of shortest paths through a node; and blocking status is TRUE if a removal of a node or link segments the network into isolated islands or components. Intuitively, connectivity and influence should have the greatest impact on the spread of a cascade fault. Generally, that is true, but if resources are limited, it might be wise to consider the most critical factors, only, so the asset can be protected.

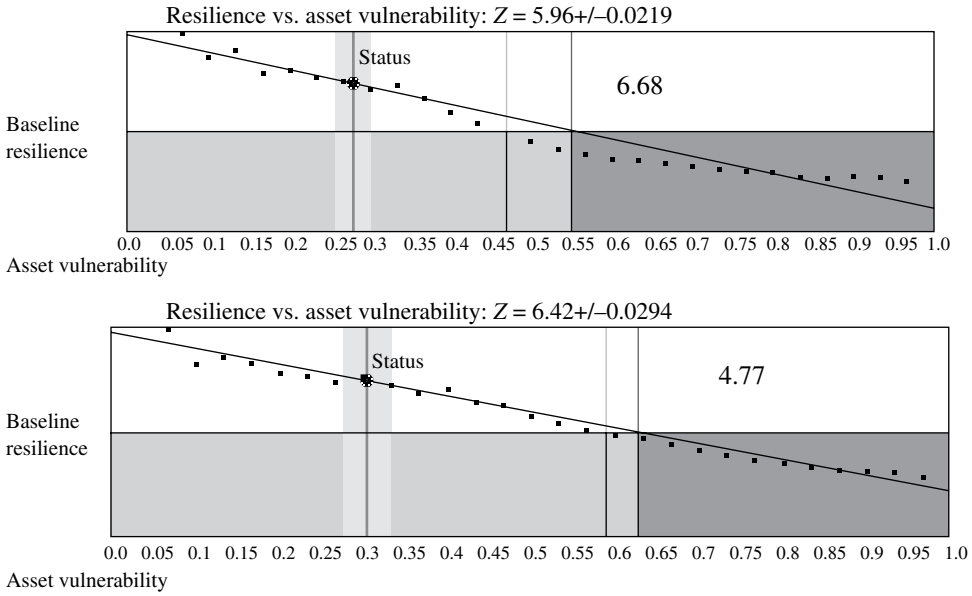


FIGURE 4.8 The normalized cascade resilience metric separates networks into high, medium, and low resilience zones as shown by shaded (colored) zones and assigns a resilience number between 0 and 10 to each. The top chart (Z = 5.96) shows the results for a network with spectral radius of 6.68, and the bottom chart (Z = 6.42) shows the results for a very similar network with spectral radius of 4.77.

Figure 4.9 contains results of simulations of cascade failures in the Los Angeles Metro commuter line containing 117 stations (nodes) and 127 links (railways). Its spectral radius is 3.69 and we assume faults spread with probability of 25%. The network diagram layout is centered on connectivity, with the most connected stations in the center.

Correlation is high between cascade frequency and connectivity (0.80), but it is even higher for influence (0.86). When only blocking nodes criticality is compared with cascade frequency, correlations are even higher—0.87 for influence and 0.89 for connectivity. Results of the simulations for the LA Metro and the Southeastern Pennsylvania Transit Authority (SEPTA) commuter railway network described next are summarized below:

LA Metro:

Connectivity: 0.80

Influence: 0.86

Blocking + connectivity: 0.89

Blocking + influence: 0.87

Betweenness: 0.37

SEPTA:

Connectivity: 0.85

Influence: 0.76

Generally, influence more accurately predicts criticality of a node during cascade collapse. This has important implications for resource allocation and protection strategies. Obviously, applying limited resources to high-influence nodes pays greater ROI than investing in low-influence nodes. Hardening of high-influence blocking nodes pays even better dividends.

- The correlation of cascade frequency versus connectivity is 0.80.
- The correlation of cascade frequency versus influence is 0.86.
- The correlation of cascade frequency versus blocking nodes and connectivity is 0.86
- The correlation of cascade frequency versus betweenness is 0.37 for nodes and −0.1 for links.

4.2.3 Targeted Attacks

The foregoing results assume faults occur randomly to any node in the network. A random number is used to select which node to attack in these simulations, but in reality, one node may be favored over another for various reasons. What happens if attacks or faults are targeted? What is the impact of targeting the most *critical* nodes on the overall risk and resilience of a CIKR network? As illustrated in the Section 4.2.2, the answer depends on the criticality of the targeted node.

Consider the simulation of major portions of the SEPTA ("Philadelphia mass transit system") serving almost 4 million people in and around Philadelphia. Figure 4.10 contains the results of simulations of cascades that could occur if one of the stations fails. A cascade in a transportation system like this represents congestion, delays, or denial of service for

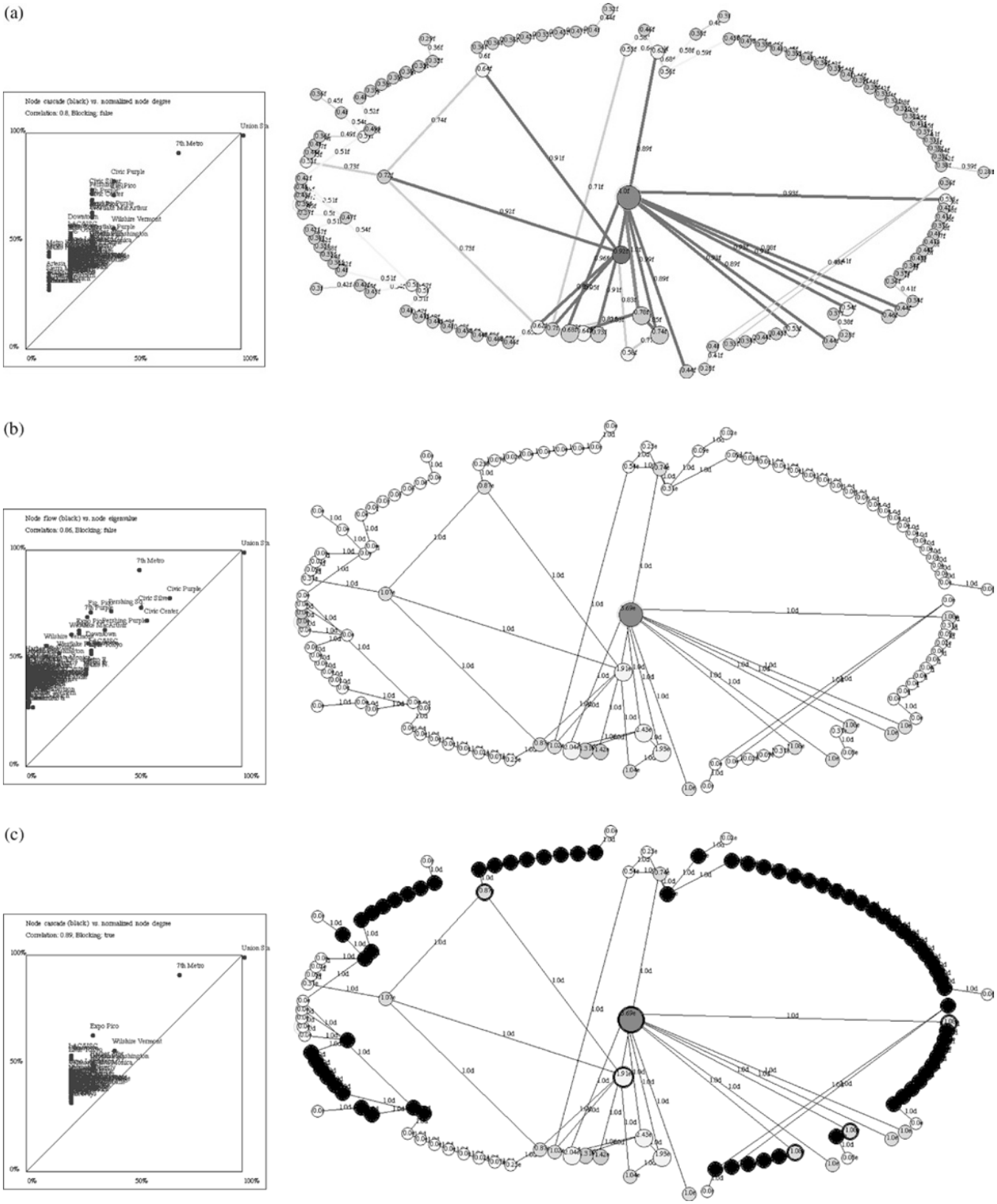


FIGURE 4.9 Heat map display results of analysis of the LA Metro railway network show high correlation coefficients between cascade frequency and connectivity, influence, and connectivity + blocking and low correlation between cascade frequency and betweenness. Correlations are slightly higher when only blocking nodes are compared with cascade frequency, indicating the exceptional criticality of blocking nodes.

(d)

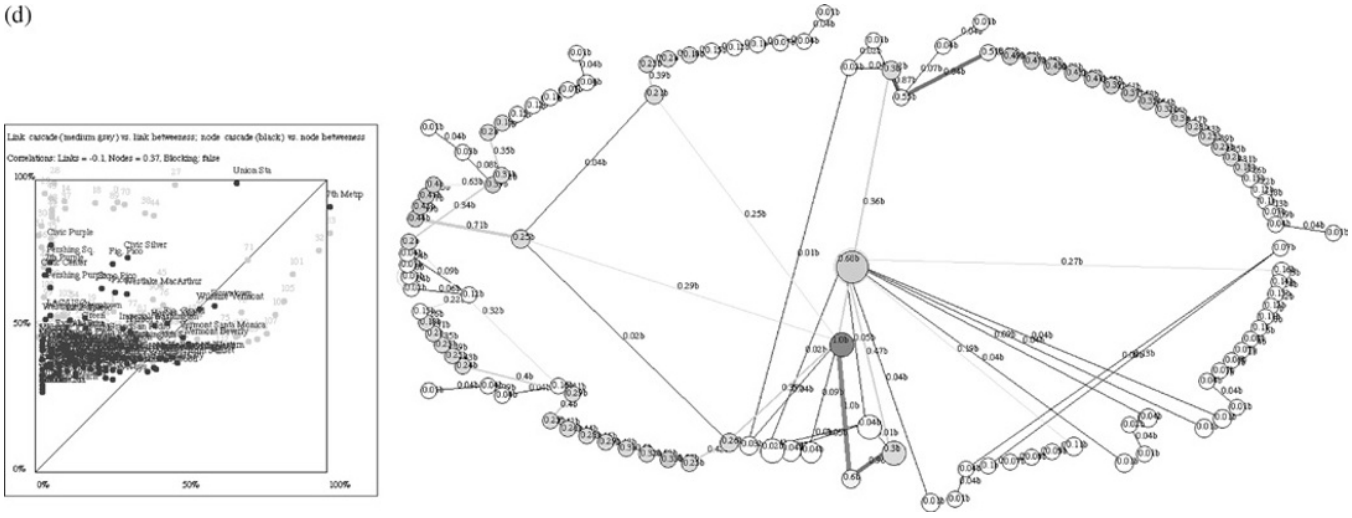


FIGURE 4.9 (Continued)

some period of time as stations go offline or trains fail to depart and arrive. If one station inserts a delay or denial of service into the network, this disturbance affects both incoming and outgoing traffic. Flow is disrupted in both directions as the disturbance spreads up and down stream.

Figure 4.10 shows the results of cascades emanating from one failed station, represented as a node. Connecting rail and bus routes are represented as links. The spectral radius of this network is 3.34 versus an average connectivity of 2.25, indicating a mild level of self-organization. In Section 4.2.2 it became obvious that high-influence blocking nodes are the most critical to cascading. The most critical nodes are high-influence blocking nodes. SEPTA contains 33 blocking nodes (37%), but in practice, it may be difficult to identify blocking nodes or calculate influence. (The author uses a software tool.)

Node connectivity of SEPTA is 86% correlated with cascade frequency, so the following analysis focuses on targeting high-connectivity nodes. Attacking highly connected nodes makes intuitive sense, because the more links a node has, the more disturbance it spreads. Hubs are *super-spreaders* simply because they have more links. From the attacker's point of view, it makes sense to target the super-spreaders. From the defender's point of view, it makes sense to allocate resources to harden super-spreaders.

Table 4.3 summarizes the results of simulations for vulnerability of 30%. Recall that high fractal dimension is better because it implies lower risk and higher resilience (see Table 4.1). In rank order of criticality, the hub at the 8th Street station is number 1; Darby station lies on the largest betweener link (Darby–Curtis Park) with 1898 paths, North Broad station is the most clustered, and 15th Street station is the furthest from the end of the line. The largest betweener links connect the Amtrak station to the International Airport:

Airport → Darby → Curtis Park → Sharon Hills → Folcroft ⋯ → Amtrak to Newark

The Airport → Amtrak chain of links is the critical path through the network because the highest betweener links and nodes lie on this chain. Even without knowing its ridership, an analyst can determine its criticality. Recall that removal of a betweener does more than anything else to separate the network into isolated components. Removing the Darby–Curtis Park link separates SEPTA into two isolated components, and removal of the second most betweener (8th Street station) divides the network into many more isolated components.

This simulation points out several general factors affecting CIKR networks. First, more links means less resilience to cascades due to *percolation*. More links connecting a hub to the other nodes exacerbates fragility, because these highly connected nodes become super-spreaders. Therefore, degree of the hub matters—the higher the degree, the more risk to the network. Higher-degreed hubs elevate the network's overall spectral radius, and when targeted, they elevate risk even more. Additionally, resilience and fractal dimension move in unison, so resilience falls as fractal dimension falls. In this example, a targeted attack on the 8th Street hub lowers the fractal dimension of SEPTA from 1.48 to 0.94, indicating a decline in resilience.

Second, structure in the form of betweenness, clustering, and highest node[6] criticality introduces far less risk than does high connectivity and influence. In this case, targeting the hub converts the power law into a cumulative binomial distribution as shown in Figure 4.10b. The frequency distribution of Figure 4.10a is centered near a consequence of 6% and is no longer long-tailed. Targeting the hub transforms normal accidents into complex catastrophes.

[6]Height is the number of hops from a node to the farthest destination or source node away.

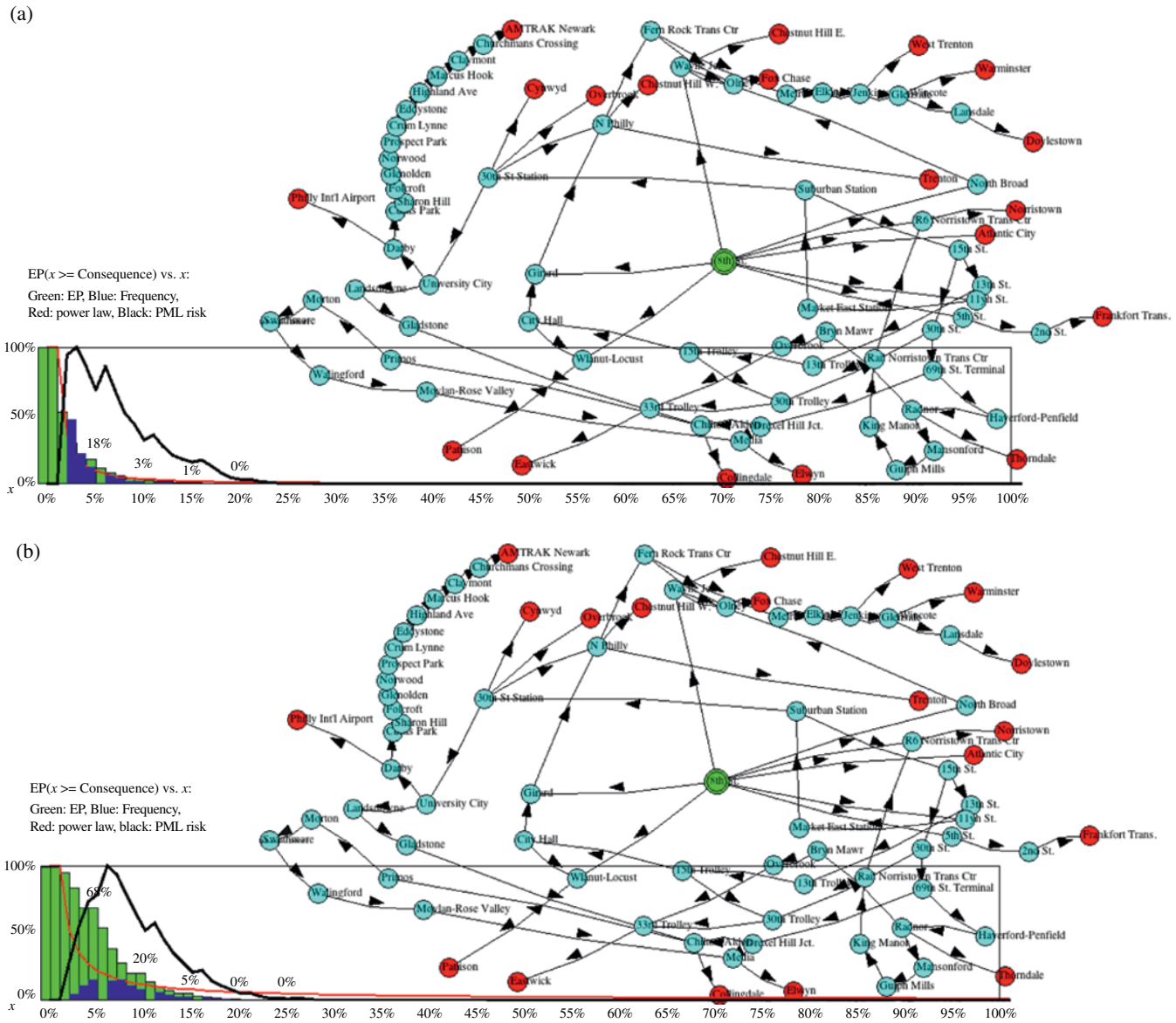


FIGURE 4.10 Risk dramatically increases when critical nodes are attacked in this simulation of targeted attacks on SEPTA. Resilience and fractal dimension move in unison so that fractal dimension is a proxy for cascade resilience. (a) Random attacks produce an exceedence fractal dimension of 1.48. (b) Targeting the hub produces an exceedence fractal dimension of 0.94.

TABLE 4.3 Fractal dimension of cascades in the SEPTA network

Attack	Asset	Fractal dimension
Random node	Node	1.48
Highest node	15th Street station	1.25
Largest cluster node	North Broad Street	1.18
Largest betweener node	8th Street station	0.94
Largest-degreed node	8th Street station	0.94

Target Attack:
Targeting a high-degreed hub (or most influential node) magnifies consequences and may result in a phase transition from simple to complex catastrophe when cascading occurs.

4.3 NETWORK FLOW RISK AND RESILIENCE

Most complex CIKR systems exist to provide a commodity such as electricity, water, gas, oil, and movement of cargo. They are *flow networks* with one or more *source* and *destination* nodes. A commodity such as passengers on a train, cargo on a boat, electrons in a transmission network, or water in a pipeline

network is moved through the network of interconnected nodes, as illustrated in Figure 4.11. A fault in one of the nodes or links may disrupt the flow and perhaps even block it.

Cascade failures partially represent the consequences of flow disruptions in flow networks if congestion, blockage, and denial of service consequences are equated with the spread of a contagious fault. For example, when an accident happens on a roadway, the flow of traffic is halted, which propagates upstream to previous intersections (nodes) and roadways (links). If the accident blocks traffic going in both directions and shuts off forward-moving traffic, then the contagion propagates downstream as well. Therefore, cascading is a suitable model of many types of CIKR failures—but not all.

Risk in flow networks can be defined as the probable maximum likelihood drop in flow when a fault occurs. The drop can be computed by comparing the sum total output across all destination nodes with the total output after a failure has occurred. The PML value of the drop can be estimated from the *resilience triangle* described in Chapter 1 or by simulation. The equation for PML risk of flow consequence is given in Appendix C.

4.3.1 Braess's Paradox

Unfortunately, simulation of network flows has severe limitations due to Braess's paradox.[7] Dietrich Braess, professor of mathematics at Ruhr University, Bochum, Germany, discovered this paradox while studying traffic patterns in transportation networks. It is considered a paradox because common sense suggests that increased capacity of individual nodes and links should also lead to increased flow through the network. But the opposite effect is sometimes observed.

> ***Braess's Paradox:***
> *Adding capacity to a flow network may reduce total flow; conversely, subtracting capacity from a flow network may increase total flow.*

If this paradox exists for a CIKR flow network, then the PML risk equation could be invalid, because it could lead to negative drops in flow. That is, the difference between flow in a fully operational network versus a damaged network might be negative, suggesting that the damaged network is more capable. A negative PML risk is more of a gain, not a loss.

To see how the paradox works, consider the simple flow network in Figure 4.11. Figure 4.11a shows the fully operational network along with current and maximum flows on each link:

Current flow / maximum flow

For example, the links flowing out from the intersection node are labeled 750/1000 and 250/250. This means that 750 units of some commodity are flowing through a link with maximum capacity of 1000 units and 250 units are flowing through a link with capacity of 250 units. The total flow reaching the destination node for this scenario is 750 units.

Figure 4.11b shows the case where the link between intersection and bypass A has been damaged. Because of this damage, the distribution of flows changes—it adapts either because automobile drivers change course, a water pipeline operator switches pipes, or an Internet router finds a better path through the network. In this scenario, 1000 units reach the destination node, by pouring more units through the operating links. The damaged network delivers more flow than the undamaged network!

But of course, this is an unsatisfactory answer. A more traditional approach is to find the *critical path* through the network and use this critical path as a means of estimating risk. Unfortunately, the traditional approach does not circumvent Braess's paradox either and often gives a false sense that risk is understood. The link removed in Figure 4.11b lies on the critical path, assuming capacity is the correct metric for criticality. The maximum capacity path traverses source → intersection → bypass A → destination. Removal of any node or link along the chain, intersection → bypass A → destination, actually improves flow. Hence, critical path analysis is an inadequate substitute for network flow analysis.

An alternative solution to network flow analysis is to use betweenness as a measure of node/link criticality. A directed network as shown in Figure 4.11 requires a directed betweenness metric that counts the number of directed shortest paths through each node and link. This metric ranks all nodes and links of equal criticality because of the symmetry of the network and the fact that betweenness counts the number of shortest paths through nodes and links. But resilience improves if alternative routes are available to redirect flow around bottlenecks or outages. So we need a metric that incorporates alternate paths even when they are not the shortest.

Figure 4.11c illustrates the multipath betweenness metric that ranks nodes and links according to their criticality along multiple paths. It assumes a system will attempt to route flow around a blockage or outage such that total output flow is maximized. In this case, there are two alternative paths, so multipath betweenness ranks nodes and links accordingly. In Figure 4.11c the rankings are:

Source → intersection: 0.83
Bypass A → destination, bypass B → destination: 0.75
Intersection → bypass A, intersection → bypass B: 0.66

Obviously, source → intersection is the most critical link because it is also a blocking link. All other links should be of

[7] http://homepage.ruhr-uni-bochum.de/Dietrich.Braess/

equal criticality, but they are not. Why? Multipath betweenness gives more weight to links nearer the destination than others. Multipath betweenness is a normalized number in the [0, 1] interval so it can be used as a proxy for vulnerability. Once we know vulnerability and consequence (flows), PML risk can be calculated by simulation.

4.3.2 Flow Network Resilience

Cascade resilience measures a network's tolerance for cascading of adjacent nodes/links. Flow resilience measures a network's tolerance for rerouting around an outage and the potential overloading that results from rerouting. Obviously, if no alternative route exists, flow resilience is zero. If one or more alternatives exist, rerouting may overload one or more nodes/links. Failure of an overloaded node/link may cause additional rerouting or failures, which lead to more overloading and more subsequent failure, until there are no more nodes/links that can fail or overloading ceases. The spread of overloading due to rerouting is a type of cascading. Overloading rather than adjacency of faults causes it.

Flow resilience depends on an overflow ratio, which is the ratio of flow to maximum capacity of node/link. A ratio of 1.5 means rerouted flow exceeds capacity by 50%. A flow resilient pipeline system, road network, or communications network might be able to tolerate an overload of 50%. But it might not, which results in a subsequent failure. The measure of flow resilience shown in Figure 4.11d is obtained by plotting loss of flow versus cutoff levels of overload ratio. That is, loss increases as tolerance declines for overflow ratio in excess of 1.0. The slope of the fundamental resilience line yields a numerical value for flow resilience.

Figure 4.11d illustrates flow resilience as it applies to the simple Braess network. Flows through bypass A overload the following links and nodes:

Intersection → bypass A: overload = 10.0

Bypass A: overload = 4.0

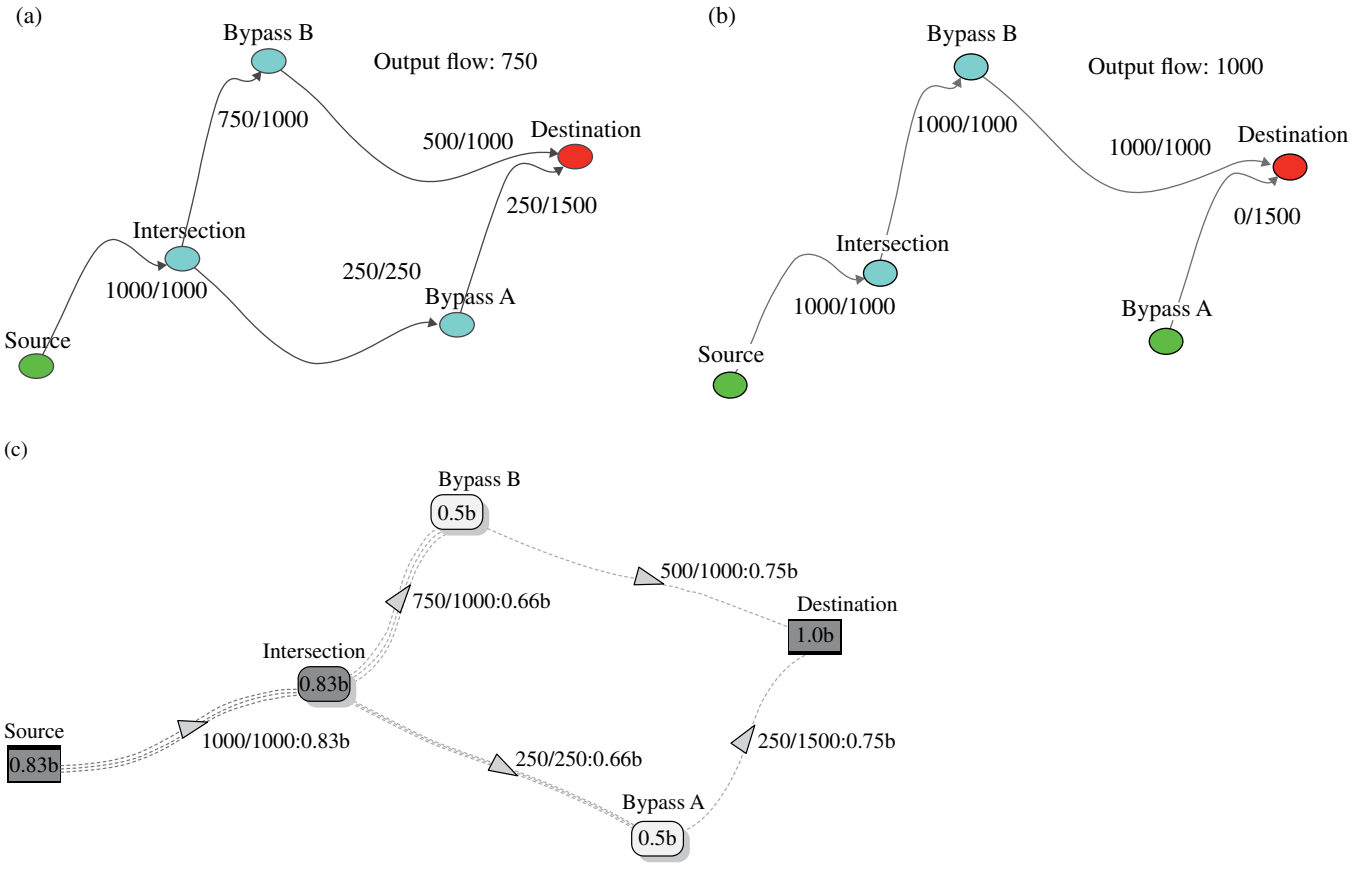


FIGURE 4.11 An illustration of Braess's paradox in network flows: removing a link increases flow from source to destination nodes instead of reducing flow. (a) Network containing 5 nodes and 5 links each marked with flow/capacity. Total flow at the destination node is 750 before the network is attacked. (b) Same network as (a) but with one link removed due to damage. Paradoxically, the total flow reaching the destination node has increased to 1000. (c) Multipath betweenness ranks nodes and links according to their criticality for resilient directional flow through a network with alternate paths between source and destination. (d) Flow overload metric and fundamental resilience line for loss of flow versus tolerance for overloading. Flow resilience of Z = 0.25 is very low.

(d)

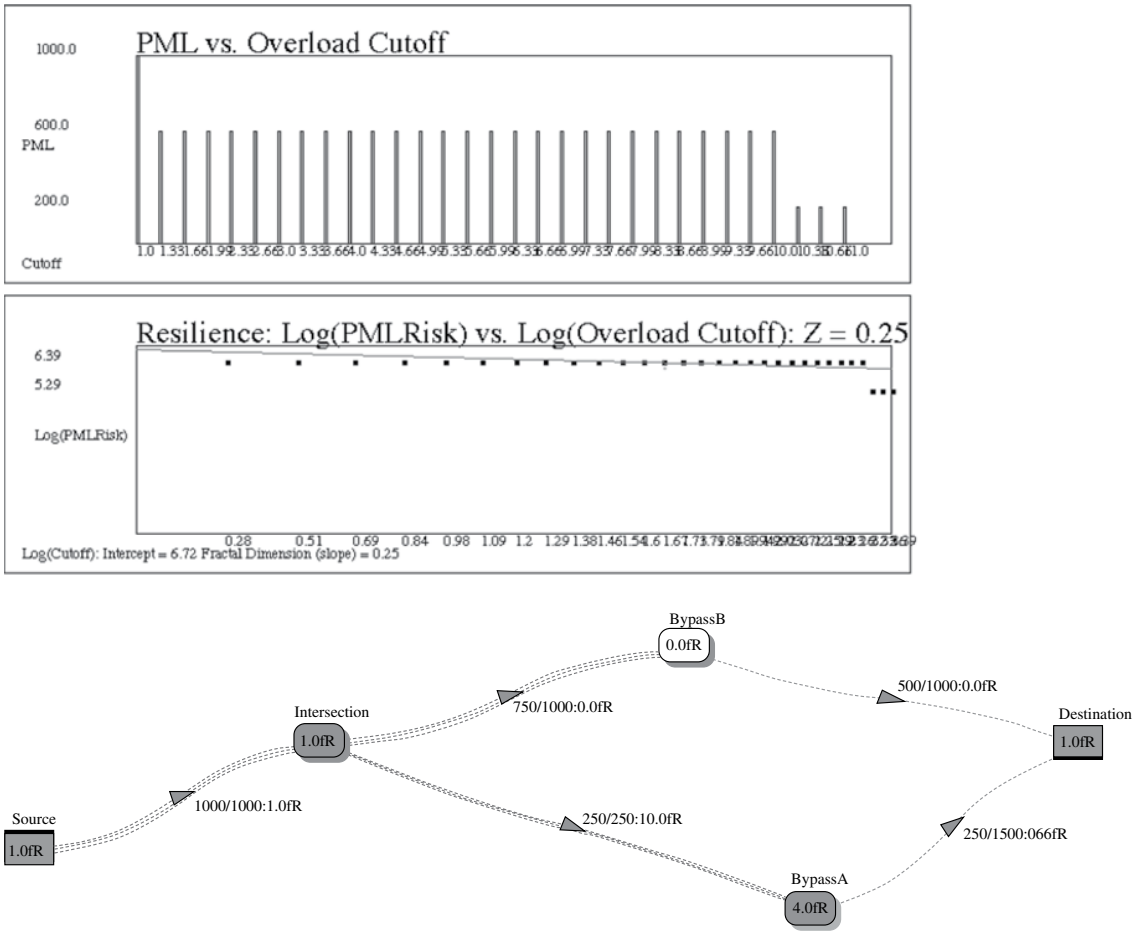


FIGURE 4.11 (Continued)

Bypass A → destination: overload = 0.66

The overload ratio of all other nodes and links is less than 1.0, so they do not fail.

The fundamental resilience line shown in Figure 4.11d plots flow PML risk versus overload threshold called the overload cutoff. Note that overload cutoff ranges from 1.0 to 8.66. PML risk declines as tolerance increases. The decline is approximated by a power law so that the slope of the resilience line.

4.4 PARADOX OF REDUNDANCY

Braess's paradox is an example not only of the paradox of enrichment (POE) but also of POR, because it says that addition of redundant components such as a duplicate traffic lane in a highway or duplicate server in an IT system may actually reduce resilience instead of increase it. More generally, adding nodes and links to a CIKR network can make it more robust against node and link failures, but it also makes it more vulnerable to cascade failures when spectral radius increases. Robustness and resilience appear to work in opposite directions in a complex CIKR network.

Redundancy may reduce single points of failure, but it may also increase risk due to cascades. For example, a redundant computer system may accelerate the spread of malware if spectral radius of vulnerability is increased as a result of adding redundancy. From a network perspective, complex CIKR systems contain two types of redundancy: link and node. Both types impact the resiliency in both positive and negative ways.

4.4.1 Link Percolation and Robustness

Link percolation is the process of adding network links. It increases SOC and therefore reduces cascade resiliency. De-percolation is the opposite process—deletion of links. It reduces SOC and therefore increases cascade resilience. But de-percolation eventually dismantles a network by separating

a connected network into disconnected components called *components* or *islands*. Link de-percolation can fragment a network such that it is impossible for a commodity such as gas, oil, or water to flow from one node to other nodes. At some point in repeated de-percolation, the network breaks up and becomes incoherent.

In terms of connectivity, link percolation provides alternative paths through a network and, therefore, flow redundancy. Generally, redundant paths make it possible for a commodity to flow through an alternate path if the primary path fails. The Internet is highly percolated so that email has many alterative paths connecting sender and receiver. If one path fails, there are plenty of other paths to take its place. In this sense the Internet is robust. But this high level of redundancy also makes the Internet extremely vulnerable to the spread of malware, because more links means more ways to spread it. The Internet suffers from the *POR* for the same reason it is robust.

Link percolation is a measure of robustness and redundancy in networks due to the number of links "holding the network together." It is equal to the number of links that can be removed without separating the network into isolated parts. A good approximation can be obtained by noting that a connected network must have at least one link for every node. Therefore, $(m - n)$ links are the most that can be removed before the network separates into islands of disconnected parts. The fraction of links that can be removed is:

$$\kappa_L = \frac{m-n}{m} = 1 - 2/\lambda$$

m: # links

n: # nodes

λ: mean degree

Figure 4.12 shows that this equation is a good approximation. For example, the top 500 airports and connecting routes shown in Table 4.2 form a network with $n = 500$ nodes (airports) and $m = 7360$ links (routes). Therefore, its link robustness is relatively high:

$$\kappa_L(\text{Airports}) = \frac{7360 - 500}{7360} = 93\%$$

This means that 93% of the routes have to be removed before separating the network into isolated islands of airports. With less than 93% of its routes removed, the global commercial air travel network remains intact. It is possible to travel from one airport to any other in the network without getting stranded. (This network assumes all routes work in both directions.) Link robustness is a form of resilience due to redundancy, but it also increases SOC, hence the POR.

4.4.2 Node Percolation and Robustness

Similarly, node percolation is a measure of robustness in networks due to the number of nodes "holding the network together." It is equal to the number of nodes that can be removed *without* separating the network into isolated components. A rough approximation is obtained by noting that node removal also deletes links that hold the network together. In a structured network, the number of deleted links per node is proportional to the spectral radius of the network. Node de-percolation removes links as a by-product of node deletion and is approximated as follows:

$$\kappa_N = 1 - 1/\rho$$

ρ spectral radius

For example, the spectral radius of the commercial air travel network of Table 4.2 is 45.3. Therefore, this network is self-organized to a very high level. It has low cascade resiliency—congestion or airport closure leads to system failure. But it is very robust with respect to node removal. In fact, 98% of the nodes can be removed (one at a time) before the network becomes separated into islands of connected airports:

$$\kappa_N(\text{Airports}) = 1 - 1/45.3 = 98\%$$

Therefore, the network formed by the busiest 500 airports is very robust. Ninety-eight percent of the airports can be removed—only one at a time—before the network separates into islands. Two percent (10 airports) are critical, however, because their removal would separate the network making it impossible to travel from any airport to any other airport.

Figure 4.13 shows very good agreement with the formula for link robustness but very poor agreement with the formula for node percolation. But they both conform to a general trend—as mean connectivity and spectral radius increase, robustness also increases. But cascade resilience decreases. Robustness works in opposition to cascade resiliency because redundancy increases risk according to the POR.

Robustness:
Robustness and cascade resiliency are inverses of one another: a robust CIKR network tends to be less resilient against cascade failures due to redundancy of nodes and links; but a robust network can tolerate more node and link removals due to redundant paths.

4.4.3 Blocking Nodes

Node robustness is a measure of how many nodes are redundant—their removal *does not* separate the network into islands. The remaining nodes are just the opposite—they are essential to holding the network together. Removal of the

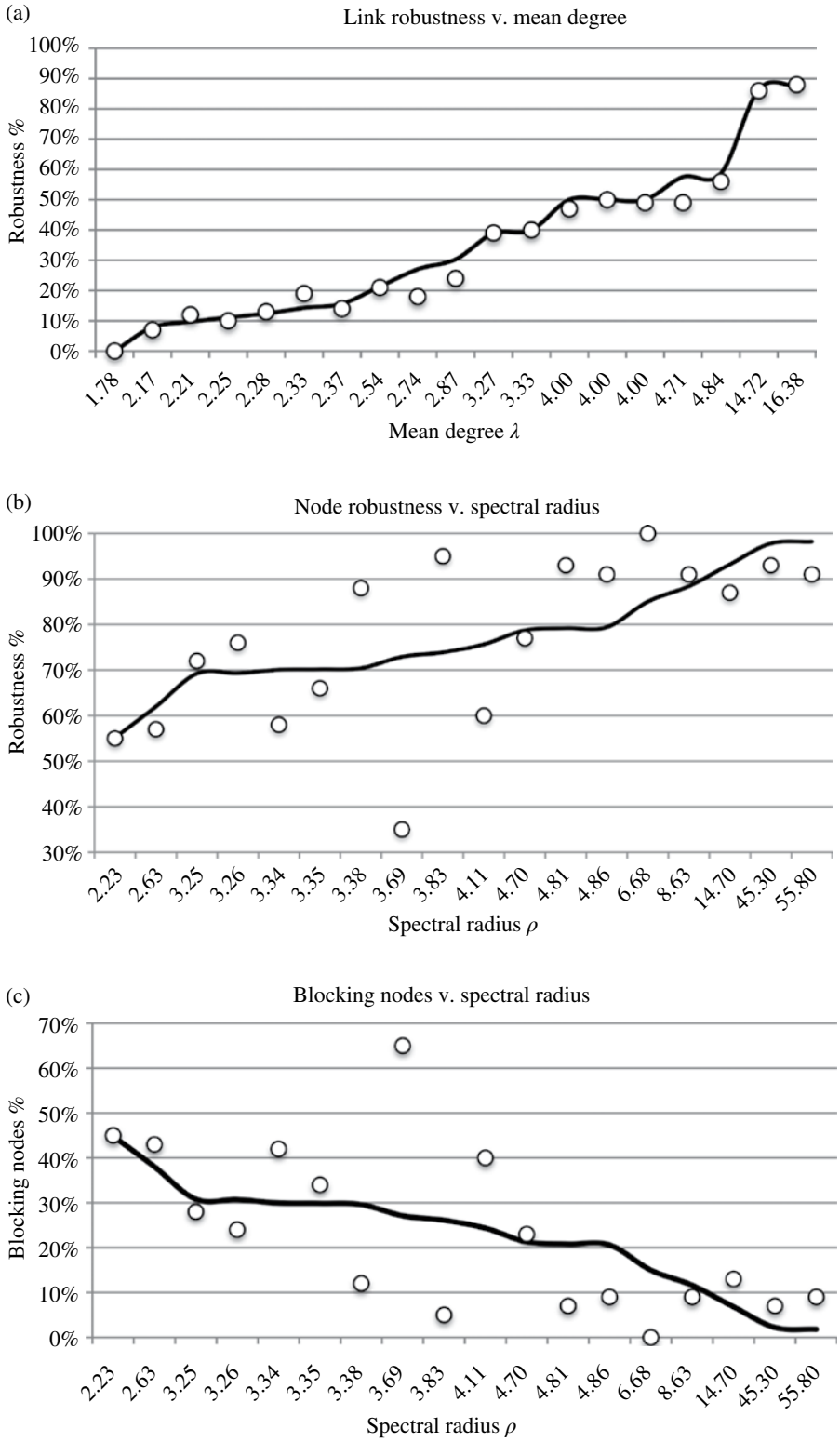


FIGURE 4.12 Robustness is a measure of the fraction of nodes and links that can be removed without separating a network into disjoint parts. The following graphs were obtained by applying link and node robustness measures to the networks of Table 4.2. (a) $1 - 2/\lambda$ is a very good approximation to link robustness. Mean degree is equal to the average connectivity per node. (b) $1 - 1/\rho$ is a very rough approximation to node robustness. Spectral radius is equal to the largest influence factor. (c) $1/\rho$ is a very rough approximation to the fraction of blocking nodes.

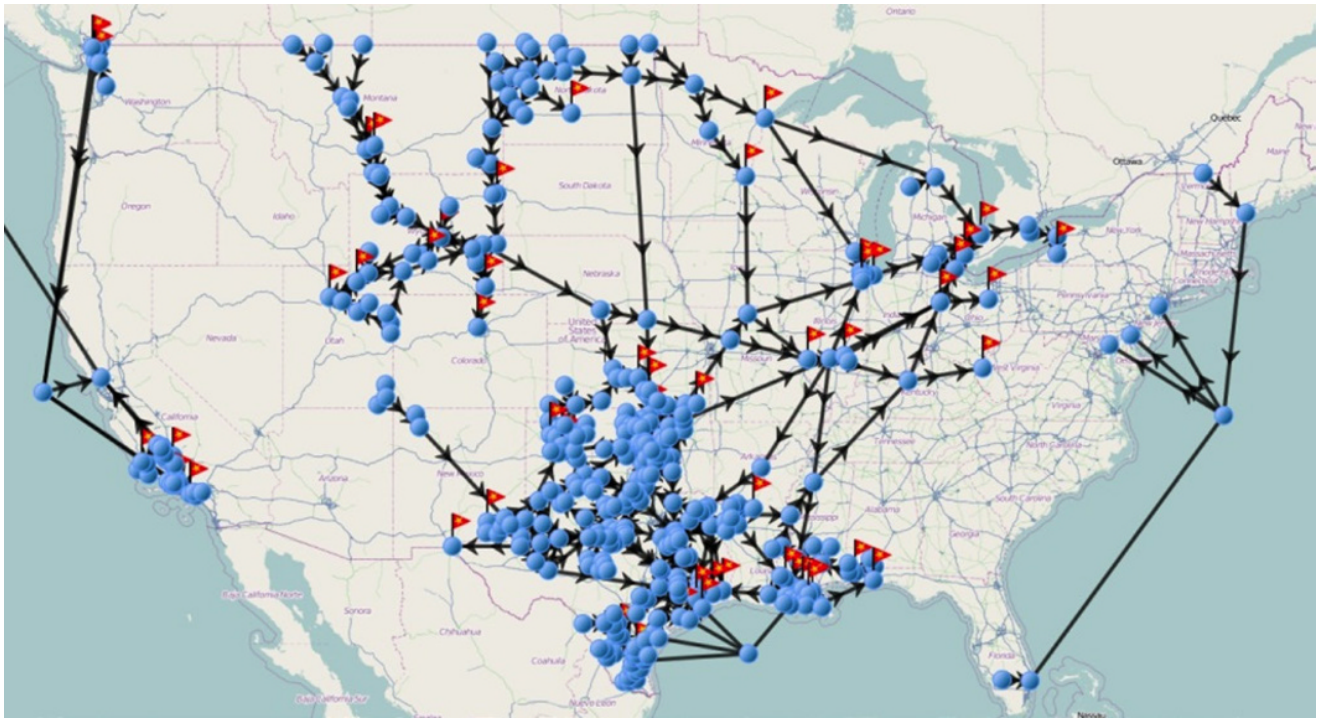


FIGURE 4.13 MBRA model of the crude oil transmission pipeline system in the United States illustrates the use of network analysis to determine risk and resilience in a complex CIKR system.

remaining $n[1-(1-1/\rho)] = n/\rho$ nodes separates the network into isolated components. These nodes are called *blocking nodes*, because without them, cascades are blocked. A blocking node is a bridge between components that would otherwise be unreachable. They are critical to the spread of a contagious disease, normal accident, or continuity of commodity flow in a pipeline or communications network. If they are removed or hardened, spreading is stopped.

There are approximately n/ρ blocking nodes in a network with n nodes. Therefore, $1/\rho$ is the fraction of nodes holding the network together. This is the blocking node fraction for a network with spectral radius ρ. For example, the network formed by the 500 busiest airports in Table 4.2 contains approximately 500/45.3 = 11 blocking nodes. Removal of all 11 of these nodes will fragment the air transportation network making it impossible to travel from one airport to all others.

Which 11 of 500 airports are the blocking nodes? The answer to this will be revealed in Chapter 14, where the best way to halt the spread of a contagious disease is through inoculation and/or quarantine of blocking nodes. By removing or protecting blocking nodes, the epidemiologist prevents a deadly contagion from spreading. Similarly, hardening of blocking nodes can prevent cascade failure of electrical power grids and pipeline networks.

Criticality of Blocking Nodes:
A CIKR network is maximally segmented into disjoint components (islands) by removal of its blocking nodes. Blocking nodes stop the spread of cascade failures across components by de-percolating the CIKR network. There are approximately n/ρ blocking nodes in a network with n nodes and ρ spectral radius.

The number of blocking nodes for each network in Table 4.2 is estimated by multiplying $1/\rho$ by the total number of nodes in each network. Identification of which nodes are blocking nodes is left as an exercise for the reader. (Hint: This is best left to a computer that examines each node one at a time to determine if its removal separates the network into disjoint components.)

4.5 NETWORK RISK

The foregoing theory of catastrophes ignores a number of other factors affecting risk. For example, it assumes uniformity of threat, vulnerability, and consequence across all nodes and links. It ignores Braess's paradox, tragedy of the commons (TOC), POE, POR, and the competitive exclusion principle. These all affect the shape and SOC of CIKR networks. For example, if one low-connectivity node represents a high vulnerability and consequence, it may be more critical than its connectivity indicates. A better model of network risk incorporates expected utility risk into each node and link so that overall network risk is more accurate. In this model, each node and link has five values:

T: probability of threat

V: probability of failure if threatened

C: consequence

P: funding gap—resource needed to eliminate V or reduce it to a minimum

R: response funding—resource needed to reduce C

This model is incorporated in the model-based risk analysis (*MBRA*) tool described in greater detail in Appendix B. MBRA extends the network model so that threat, vulnerability, and consequence of each node and link is considered in the calculation of risk. Additionally, network properties such as connectivity and betweenness may be incorporated into the calculation of risk to account for self-organizing factors that determine the architecture (topology) of the CIKR network (see Appendix B for details).

Static risk is the sum of TVC over all nodes and links. Dynamic risk is defined as the maximum of PML risk for cascades and losses of flow. A PML risk curve is easily obtained by multiplying consequence (x-axis) by EP and observing that maximum PML occurs at the maximum value of the resulting curve. Consider the following example as an illustration of these concepts.

4.5.1 Crude Oil and Keystone XL

Perhaps the best way to explain static and dynamic risk is to use them to analyze a major CIKR network. The US crude oil network became the focus of national attention over the *Keystone XL* pipeline controversy circa 2011. This major pipeline network is designed to carry Canadian shale oil to Cushing, Oklahoma, storage tanks and eventually to the vast refinery resources of the Texas gulf region (see Fig. 4.13). The question posed by Michael Larrañaga in his thesis was, "Does KeystoneXL increase the resilience of the US crude oil network?" Larrañaga and associates used MBRA to answer this important question [1].

Each node and link in Figure 4.13 is annotated with the following input values:

Name of asset

Threat probability, T

Vulnerability probability, V

Consequence, C

Consequence prevention cost (max), P

Consequence response cost (max), R

Threat, vulnerability, and consequence are used to calculate node/link risk using the PRA equation R = TVC. The prevention and response costs are used to allocate prevention and response budgets to nodes and links such that risk is minimized across the entire network. The consequence prevention cost is an estimate of the cost to reduce vulnerability to some minimum—typically 5% of the initial vulnerability. Consequence response cost is an estimate of the cost to reduce consequence to some minimum—also typically 5%. In other words, these costs are estimates of resources needed to eliminate 95% of the vulnerability and consequence values entered as input.

MBRA assumes a diminishing returns model of vulnerability and consequence reduction, so that it is impossible to eliminate all vulnerability and all consequences. Similarly, MBRA assumes an adversary has limited resources and also suffers from diminishing returns. It takes an infinite investment to drive threat probability to 100%. Appendix B has mathematical details.

The answer to the question posed by Larrañaga et al. is, "Yes, addition of the KeystoneXL pipeline reduces risk and increases resiliency." An MBRA simulation of the impact of single node failure on both pre- and post-Keystone XL networks shows that fractal dimension changes from 1.17 to 1.40, an improvement in EP—evaluated at 9% consequence—of 74%. PML risk at the highest consequence in Figure 4.14 (9%) is reduced by 55%.[8] Keystone XL makes a big difference to cascade resilience.

In addition to the reduction in risk to the nation's crude oil supply chain, the MBRA analysis identified critical components in the crude oil supply chain. These consist of critical ocean shipping routes, storage tanks, refineries, and pipeline assets. Larrañaga et al. note that the impact of even a short disruption can have significant economic consequences ranging from $97 million (3-day) to $402 million (10-day) and upward of $20 billion in secondary impacts as in the BP/Macondo/Deepwater Horizon incident.

4.5.2 MBRA Network Resource Allocation

The main purpose of MBRA, besides calculating network risk, is to optimally allocate limited resources to minimize overall risk. MBRA takes a portfolio approach and uses an iterative Stackelberg competition algorithm to iterate between reducing risk by reducing vulnerability and consequence versus increasing threat and risk. If an equilibrium point exists, MBRA finds a balance between maximum threat and minimum risk and as a by-product computes corresponding allocations of prevention, response, and attacker budgets.

It is not necessary to estimate T and V as in traditional PRA, but of course allocations are better if better inputs are used. In this example, all threats are set to 50% (maximum ignorance), and all vulnerabilities are set to 100% (worst case). MBRA will recalculate threats and vulnerabilities such that risk is minimized by the defender and maximized by the attacker. This rational actor model assumes an evil

[8] $74\% = [0.09^{-1.40} - 0.09^{-1.17}]/0.09^{-1.17} = [0.09^{1.17-1.40} - 1]; 55\% = (80\% - 36\%)/80\%$.

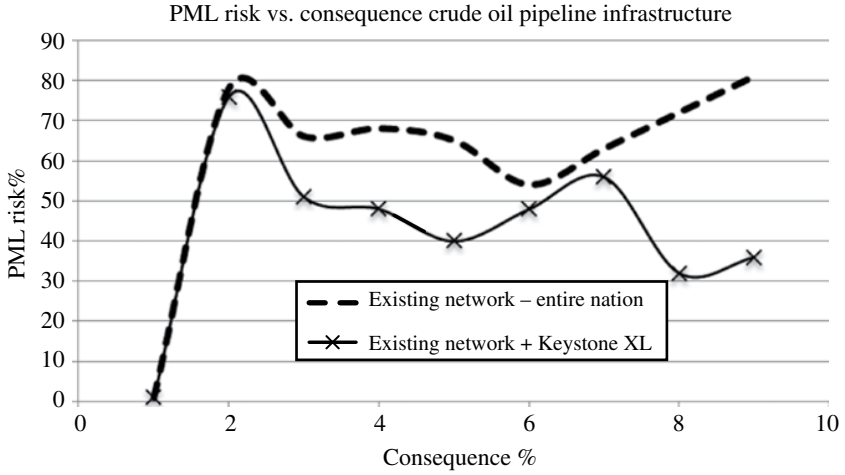


FIGURE 4.14 PML risk profiles for the crude oil CIKR network of Figure 4.13 show a significant reduction of risk after the Keystone XL pipeline is added, especially as consequence rises.

adversary always tries to maximize risk and ignores the possibility of an opportunistic attacker.

To make the example of Figure 4.15 more interesting, slightly different consequence, prevention, and response costs are input to the flow network of Figure 4.11. These inputs are shown in Figure 4.16 along with the initial risk ($6125) and risk ranking. Link names are created by combining names of the node pairs they connect. For example, Src–Inter is the name given to the link connecting source → intersection, and A–destination is the name given to the link connecting bypass A to destination. Note that destination is ranked highest in initial risk and intersection and bypass B are number two and three, respectively.

Since this is a flow network prone to Braess's paradox, we use connectivity and betweenness weights in the risk calculation. The idea is to account for both cascade failures and flow failures. Recall that normalized connectivity and normalized betweenness are multiplied together and used as a weighting factor in the calculation of network risk. The factors that influence risk are threat, vulnerability, consequence, connectivity, and betweenness. Luckily, this is an easy calculation for a computer, but not so easy for a human.

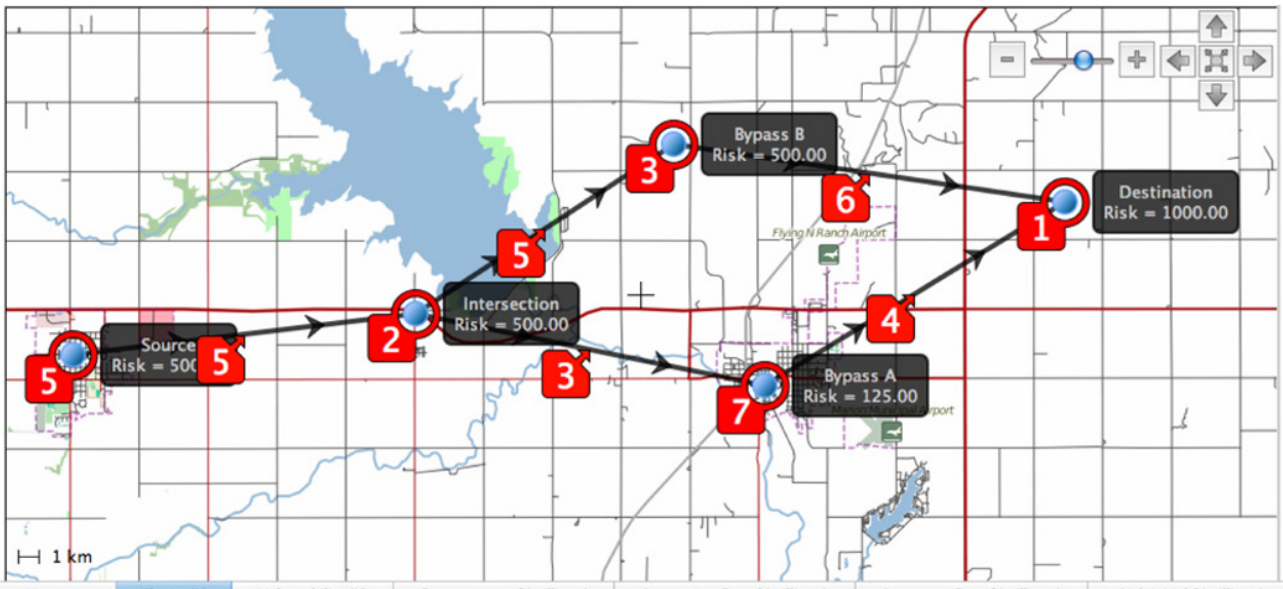
How much should be invested in prevention and response by the defender, and how much in threat by the attacker? Keeping in mind that risk reduction is susceptible to diminishing returns in this network, we can consult the diminishing return curves shown in Figure 4.16 to find a "sweet spot" for prevention and response budgets. Prevention allocation reduces vulnerability, which reduces risk; and response allocation reduces consequence, which also reduces risk. From Figure 4.16 it is determined that the risk reduction (28% of $6125) drops off after $3900 is invested in each of the defender's prevention and response budgets and $3900 (82%) for the attacker's threat budget. Using these values, MBRA allocates resources to nodes and links and calculates a reduced risk of $1175. This produces an ROI of (6125 − 1175)/7800 = $0.63/$. Unfortunately, only 63 cents or risk reduction is obtained for each dollar invested.

If $1950 is invested in prevention and response, the ROI is $1.06/$. This is still a relatively low return. An ROI of $1.11/$ is obtained for investments of $1500 each. Suppose this is a satisfactory return. What are the optimal allocations to nodes and links using these modest investment amounts? The results of $1500, $1500, and $3900 each for prevention, response, and attacker threat are summarized in Figure 4.17. The top three assets are bypass B, intersection, and the links: Inter–B and Src–Inter. These are allocated resources by MBRA to reduce vulnerability and consequence and adjust threat probabilities (either up or down), such that risk is minimized from the defender's point of view and maximized from the attacker's point of view.

Note that zero investments are made to the two links. Why? Their prevention and response costs are the highest ($5000) of all nodes and links. MBRA determined that they are a poor investment because ROI is so low. MBRA favors the highest ROI nodes and links over lower ones.

Also note that source node and Inter–A link are high-risk assets after allocation, and yet they ranked 5th and 6th out of the 10 assets in criticality. Why? Their normalized degree and betweenness values were the lowest at 0.19, compared with 1.0 and 0.48 for the top two nodes. Risk is multiplied by these weights to obtain the ranking. (In a flow network, the source node is all important, because it is the source of the flow! To analyze the importance of the source node in MBRA, use the height–weight instead of betweenness.)

Resource allocation reduces network risk and increases resiliency enough to transform this network from a complex



Name	Threat (%)	Vulnerability (%)	Consequence $(millions)	Prevention Cost $(millions)	Response Cost $(millions)	Risk Initial $(millions)
Bypass B	50.00	100.00	1000.00	3000.00	3000.00	500.00
Intersection	50.00	100.00	1000.00	2000.00	2000.00	500.00
Src-Inter	50.00	100.00	1000.00	5000.00	5000.00	500.00
Inter-B	50.00	100.00	1000.00	5000.00	5000.00	500.00
Destination	50.00	100.00	2000.00	1000.00	1000.00	1000.00
Source	50.00	100.00	1000.00	1000.00	1000.00	500.00
Inter-A	50.00	100.00	2500.00	1000.00	1000.00	1250.00
B-Destination	50.00	100.00	1000.00	500.00	500.00	500.00
A-Destination	50.00	100.00	1500.00	500.00	500.00	750.00
Bypass A	50.00	100.00	250.00	500.00	500.00	125.00

FIGURE 4.15 MBRA model of the flow network of Figure 4.11 layered on a map, showing input values and initial risk.

catastrophe prone network to a normal accident. The average TV value is reduced from 50 to 42% after allocation. But the simple network is not very robust—link robustness is 0%, node robustness is 53%, and so its blocking node fraction is 47%. Removal of one link is enough to separate the network into disjoint islands, and removal of slightly less than one-half of the nodes also leads to network separation. Which link and which nodes? This is left as an exercise for the reader.

Overall network risk drops from $6125 to $2792 with an ROI of $1.11/$. This means risk is reduced by $1.11 for each dollar invested. But an interesting thing happens when the attacker increases the threat budget. A critical point is reached at an attacker budget of $7098–$7099. Risk unexpectedly jumps from $3233 to $4630 as the attacker invests $1 more than $7098. Why?

Table 4.4 shows how the MBRA optimizer shifts attacker investments when this critical point is reached. When the attacker budget increases from $7098 to $7099, the attacker can afford to threaten two nodes instead of only Bypass B. This is shown in Table 4.4 as a shift in resources by both attacker and defender. The ROI of the defender drops to $0.50/$. The attacker essentially outspends the defender.

Why does this nonlinear change in risk happen? Diminishing returns explains this phenomenon, too. MBRA approximates diminishing returns as an exponential curve like the ones shown in Figure 4.16. The lopsided curves diminish rapidly at first and then drop off more slowly as investment increases. The MBRA optimizer "rides the steepest curve" until its rate of decline tapers off and then switches to another node/link with a more rapid rate of decline when the optimizer finds a steeper path to the minimum risk. The rates of decline are determined by the ratio of prevention, response, and threat costs versus consequence. Every node/link rate of decline is typically different. The abrupt change in risk versus investment can be traced to these differing rates.

The steps in performing an MBRA analysis are as follows:

1. Build a network model of the CIKR system and enter estimates of T, V, C, prevention, and response costs for every node and link.

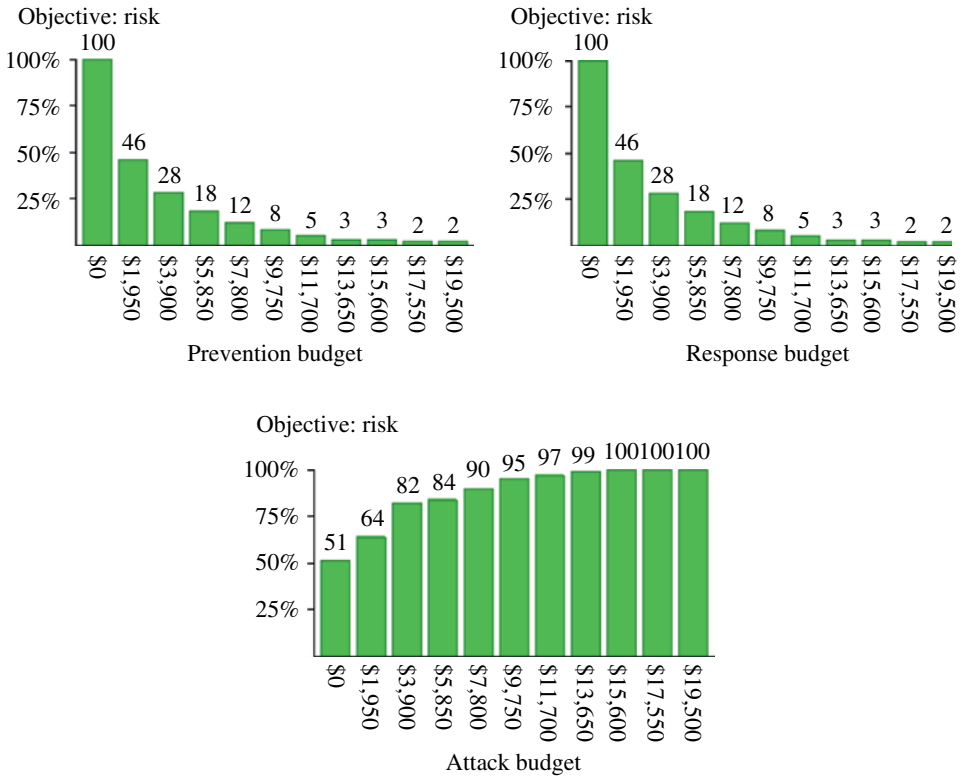
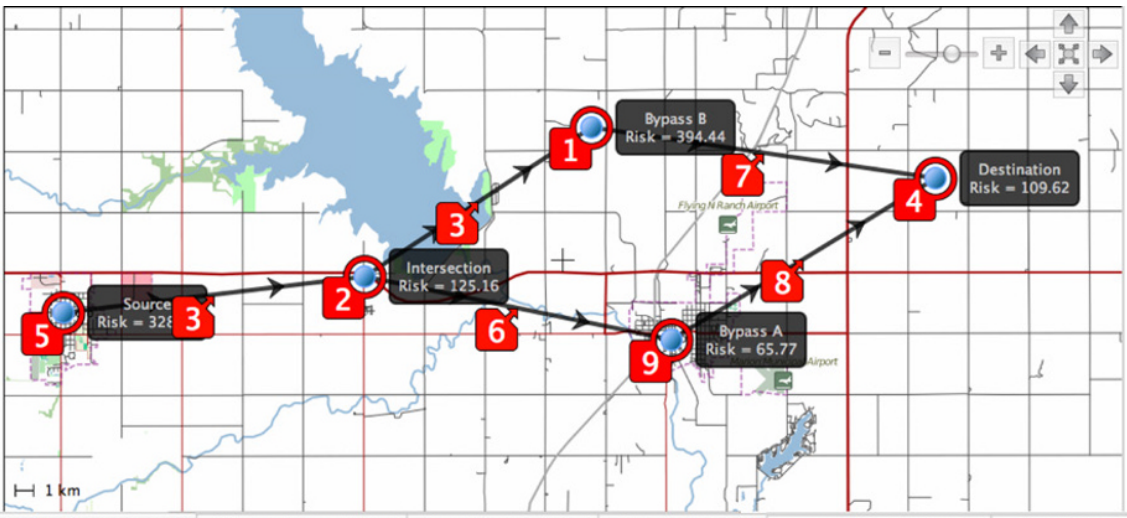


FIGURE 4.16 The diminishing returns curves for prevention, response, and threat budgets suggests optimal return on investment at $3900 (28%) for prevention and response and $3900 (82%) for optimal return on investment by the attacker.



Name	Prevention Allocation $(millions)	Response Allocation $(millions)	Attack Allocation $(millions)	Vulnerability Reduced (%)	Consequence Reduced $(millions)	Calculated Threat (%)	Weight
Bypass B	249.44	249.44	1228.73	68.19	681.87	84.83	0.476
Intersection	415.67	415.67	820.37	38.40	384.00	84.88	1.000
Src-Inter	0.00	0.00	0.00	100.00	1000.00	50.00	0.190
Inter-B	0.00	0.00	0.00	100.00	1000.00	50.00	0.190
Destination	297.54	297.54	410.81	25.41	508.10	84.92	0.571
Source	102.95	102.95	409.76	62.24	622.45	84.85	0.190
Inter-A	202.50	202.50	410.25	39.35	983.86	84.88	0.190
B-Destinatior	73.35	73.35	205.50	50.89	508.87	84.93	0.143
A-Destinatio	95.50	95.50	205.20	41.50	622.44	84.89	0.143
Bypass A	63.61	63.61	205.33	55.66	139.16	84.91	0.476

FIGURE 4.17 Risk ranking and calculated results after allocation of $1500 to prevention, $1500 to response, and $3900 to threat.

TABLE 4.4 Change in allocation is abrupt when the attacker budget exceeds $7098

Attacker budget	$7098	Prevention	Response	Threat
	Destination	$168	$168	$0.00
	Bypass B	$0.00	$0.00	$1371
Attacker budget	$7099			
	Destination	$0.00	$0.00	$1000
	Bypass B	$354	$354	$1099

2. Decide which network properties to use, for example, degree, betweenness, and so on.
3. Find an optimal ROI point by trial and error. MBRA provides risk-versus-investment tools as illustrated in Figure 4.16. It may take several attempts to find an overall ROI in excess of $1.00/$.
4. Increase the attacker budget to explore the possibility of a critical point. Revisit the ROI question: is a better ROI possible at a critical point?
5. Simulate cascade failures on the network before and after allocation to obtain fractal dimensions. Compare the fractal dimensions to determine the effectiveness of risk reduction. Is the network prone to low-risk, high-risk, or complex catastrophes?
6. Further exploration of "what-if" scenarios involving changes in input values may provide additional insights. Are the results sensitive to small changes in inputs?
7. Make recommendations: what is the best use of resources, and over what period of time? Should the investment be made all at once or spread out over time?

4.6 THE FRAGILITY FRAMEWORK

Hodges proposed a qualitative framework for evaluation of critical infrastructure whole-of-community resilience called the Fragility Framework [2]. Whole-of-community resilience is defined in terms of the people and social systems needed to prevent CIKR failures, respond to threats and incidents, and recover from threats and incidents through community action. In the Hodges framework, resilience has a causal relationship with complex CIKR system dimensions such as stability and sustainability, inclusive of the organizations responsible for CIKR as well as the physical and virtual structures themselves.

People and organizations are the main elements of Hodges' framework. CIKR functions in a community are defined as "Unified groups of individuals with a common purpose before, during and after emergencies and disasters." Furthermore, civil society operates within an assumed "whole community" environment where fire, law enforcement, emergency responders, public health, and various government agencies work together to protect human lives as well as physical and virtual infrastructure.

The Hodges framework models CIKR community resilience in terms of *system fragility*, which is defined as "A quality that leads to weakness or failure within a system, sometimes resulting in cascading effects (the domino) that can lead to systemic failures and collapse." Furthermore, *community fragility* is defined as "A quality that leads to weakness and possible failure within a system of systems which connects emergency management to an affected community." Fragility and resilience are inverses of one another, although the Hodges framework more narrowly defines resilience in terms of recovery, "The ability of a system to return to its original form or position, or the ability to recover quickly." This narrow definition ignores the contribution of prevention and adaptability to resilience.

Resilience is often called antifragility. These two terms will be used interchangeably, except where meaning might be obscured. Additionally, it is useful to quantify resilience—aka antifragility—in numerical terms so one community can be compared with another and to make objective resource allocation decisions. The generally accepted definition of resilience is "the ability to resist, absorb, recover from or successfully adapt to adversity or a change in conditions" (see Appendix F). The Hodges framework is flexible enough to incorporate this broader definition as well as additional complexity measures described in this book. Therefore, the Hodges framework can also be expressed in terms of a fault tree, which has the added benefit of optimally allocating resources to reduce fragility. An extension of the Hodges framework to fault tree analysis and resource allocation is detailed below.

4.6.1 The Hodges Fragility Framework

Figure 4.18 summarizes the general framework proposed by Hodges in 2015 during the whole-of-community era of the Department of Homeland Security. The framework consists of three dimensions: connectedness, stability, and sustainability, although one can easily extend these dimensions. Each dimension has a collection of causality factors that drive fragility along a dimension. A causality factor connects a process or state (the cause) with a dimension (the effect), where the first is partly responsible for the second and the second is partly dependent on the first. In general, there are many causes of fragility. Inversely, resilience is improved by the removal or mitigation of causal factors.

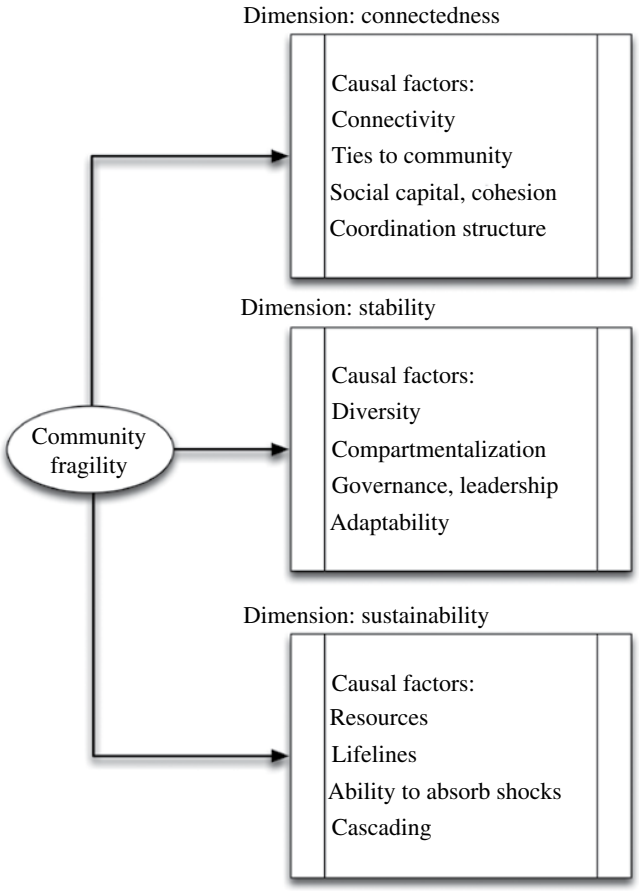


FIGURE 4.18 The Hodges Fragility Conceptual Framework defines fragility in terms of the complex interrelationships in the community responsible for response to CIKR incidents. Hodges identifies three dimensions (connectedness, stability, and sustainability) and 12 causal factors.

Connectedness means connectivity between emergency management services and the community it serves. This includes the development of social capital, strong connections with community organizations such as Red Cross, churches, schools, and government agencies that must come together during and after an event.

Stability is most impacted by strong leadership within the community, the flexibility of emergency response plans, and the degree of compartmentalization in emergency management systems. Weak leadership can lead to a lack of trust, further affecting social capital. On the other hand, strong leadership can lead to better planning efforts with more flexibility and adaptability. Moreover, compartmentalization of the emergency management system leads to greater stabilization and less of a chance for small disturbances to cause major problems.

Sustainability includes CIKR properties such as building resiliency, resource availability, lifeline/backup supplies, and the identification of potentially cascading events. For example, food and water are lifeline resources. Food and water backup/restoration and supply chain management allows a community to recover faster and more robustly compared with an unprepared community that has no backup supplies.

The Hodges framework has been applied to several natural disasters—fires in California, the Katrina Hurricane in New Orleans, and a major tornado in the Midwest. Hodges scored fragility in terms of qualitative measures: low, medium, and high fragility. This makes its application easy to use, but does not allow decision-makers to optimally allocate resources to different regions of the country based on quantitative metrics. Therefore, the author proposes a simple scoring mechanism that yields numerical scores based on estimates of antifragility or resilience.

Each causality factor is rated on a resilience scale of 0–100. Dividing out the maximum total points normalizes the sums across the causal factors along each dimension. In this case, the maximum total is 1200 points. The sum across dimensions is assigned the overall resilience for the community. Consider the example shown in Table 4.5. Each causal factor is rated on a scale of 0–100. The scores are summed and normalized by dividing by 1200. The normalized sums are totaled to yield an overall score.

The worksheet of Table 4.5 quantifies levels of resilience on a 0–1.0 scale for each dimension. Connectedness is the most resilient, while sustainability is the least resilient. Overall community resilience is 0.496 or about in the middle of the scale. A perfect score of 1.0 is unlikely to ever occur in practice, but communities should strive for a score of 0.75 or more in general.

4.6.2 The Hodges Fault Tree

The Hodges framework is flexible enough to be adapted to different sectors and different dimensions. For example, the dimensions proposed in the chapter on theories of catastrophe combined with network science metrics for risk and resilience may be used in a general framework aimed at combining people and organizations with physical and virtual assets in a CIKR. The reader may recall them:

- Connectedness and spectral radius: 0–100
- Stability and POE: 0–100
- Sustainability and TOC: 0–100

These added dimensions might be used to enhance the community/people/organizational dimensions of the Hodges framework. More importantly, however, the Hodges framework analysis is isomorphic with fault tree analysis. The root of the fault tree is community fragility as in the qualitative framework of Figure 4.18. Dimensions are components, and causality factors are threats in the corresponding fault tree model. A generic

TABLE 4.5 Example of scoring the Hodges Fragility Framework

Dimension	Factor	Factor	Factor	Factor	Sum of scores
Connectedness	Connectivity: 50	Community ties: 80	Social capital: 100	Coordination: 25	0.213
Stability	Diversity: 80	Compartmentalization: 30	Leadership: 10	Adaptability: 90	0.175
Sustainability	Resources: 50	Lifelines: 20	Absorb shocks: 50	Cascading: 10	0.108
Total					0.496

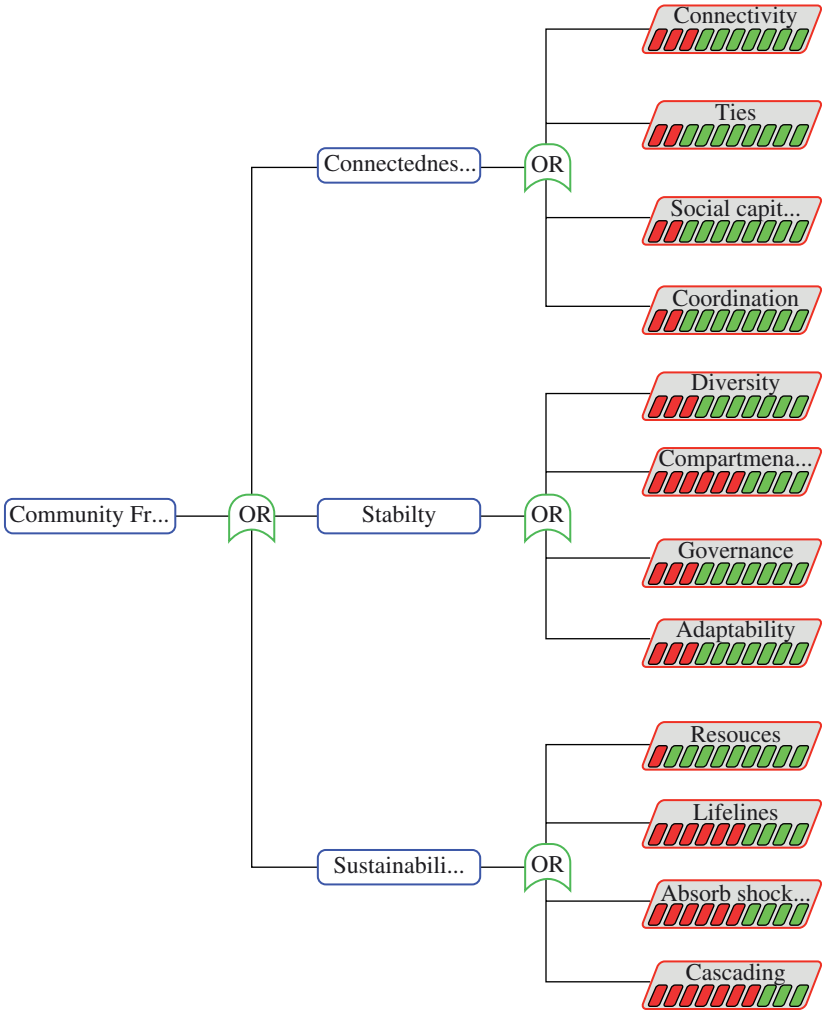


FIGURE 4.19 The Hodges framework is isomorphic to a fault tree where dimensions are components and causal factors are threats. The fault tree shown here is optimized to minimize fragility (risk of failure) by allocating $15 million across threats.

Hodges fault tree corresponding to the Hodges framework of Figure 4.18 is shown in Figure 4.19.

Fault tree analysis of the Hodges framework requires additional numerical data for risk (threat, vulnerability, consequence) and funding to reduce gaps (elimination cost). Once these are known, it becomes a simple matter of optimally allocating a budget to reduce fragility, which is quantified as risk due to failure in the fault tree. Figure 4.19 shows the results of allocating $15 million of a total gap of $43.5 million needed to remove all causal factors. Initially, risk is $269.4 million. This is reduced to $59.65 million by optimally applying $15 million across all threats in proportion to their ROI.

4.7 EXERCISES

1. In a complex CIKR network, node connectivity (degree) is defined as the:
   a. Number of links it has
   b. Size of the hub
   c. Amount of self-organized criticality
   d. Measure of self-organized criticality
   e. Amount of preferential attachment
2. Betweenness is obtained by counting the:
   a. Number of links
   b. Size of the hub

c. Degree of the hub
d. Number of paths
e. Number of shortest paths

3. Which of the following networks with n nodes and m links typically has the most SOC?
   a. Random network
   b. Scale-free network
   c. The crude oil network
   d. Clustered network
   e. Long-tailed power law
4. Self-organized criticality is increased in networks with:
   a. A lot of nodes
   b. A lot of clusters
   c. A lot of links
   d. Small diameters
   e. Large diameters
5. Emergence by rewiring or percolating a network typically:
   a. Increases structure and/or organization
   b. Increases betweenness
   c. Decreases betweenness
   d. Increases randomness
   e. Decreases structure and/or organization
6. In this book, cascading is a form of:
   a. Domino chain reaction
   b. Contagion
   c. Fault propagation
   d. Normal accident
   e. All of the above
7. A CIKR network is much more likely to suffer a complex catastrophe when:
   a. Global warming increases the severity of storms.
   b. Super storms hit a CIKR network.
   c. Vulnerability * spectral radius $\gg 1$.
   d. Vulnerability * spectral radius < 1.
   e. Spectral radius is greater than 1.0.
8. The fundamental resilience line between fractal dimension and spectral radius assumes:
   a. Uniformly random network structure
   b. Scale-free network structure
   c. Clustered network structure
   d. High spectral radius
   e. Random attacks or faults
9. The damaging effects of cascades are magnified by:
   a. Targeted attacks or faults
   b. Random attacks or faults
   c. Punctuated equilibrium
   d. Normal accidents
   e. Flow networks
10. A flow network may actually perform better when partially damaged due to:
    a. A critical point
    b. Percolation
    c. Braess's paradox
    d. SOC
    e. TOC
11. The main difference between PRA and network risk is:
    a. Network criticality factor weights
    b. Exceedence probability
    c. SOC
    d. Competitive exclusion principle
    e. Catastrophe potential
12. The Stackelberg competition algorithm is used by MBRA to:
    a. Calculate network risk
    b. Maximize risk due to threat and minimize risk due to vulnerability
    c. Minimize risk due to an attack and maximize risk due to defense
    d. Iterate between prevention and response
    e. To analyze critical points
13. The dimensions of the Fragility Framework of Hodges are:
    a. Consequence and time
    b. Consequence, time, recovery, and complexity
    c. Connectivity, sustainability, stability
    d. Connectivity, influence, and betweenness
    e. Risk and resilience
14. MBRA analysis of the Keystone XL pipeline showed an improvement in crude oil transmission because:
    a. Fractal dimension is increased by addition of the pipeline.
    b. Fractal dimension is decreased by addition of the pipeline.
    c. Spectral radius is greater.
    d. Network diameter is greater.
    e. Node and link infectiousness is lower.
15. Increasing a network's spectral radius usually increases its:
    a. SOC
    b. TOC
    c. POE
    d. Competitive exclusion
    e. Stability

4.8 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. PML risk and resilience are related through fractal dimension. Explain the relationship. Hint: Start with Table 4.1.
B. What are the pros and cons of a detailed network science analysis versus the less quantitative Hodges Fragility Framework?
C. There is a causal relationship between cascade frequency of a node and its influence. How does influence

(or connectivity) impact your strategy for infrastructure protection, or does it?

D. There is a causal relationship between overloading of a node or link and disruptions due to loss of flow in a flow network. How does betweenness and overloading factors impact your strategy for infrastructure protection, or does it?

E. The resilience of the Keystone pipeline network was improved by addition of the XL portion of the pipeline. This is a form of percolation. What is percolation, and how can it improve flow resilience when it also tends to increase spectral radius?

REFERENCES

[1] Smith, P.K., Bennett, J.M., Darken, R.P., Lewis, T.G. and Larranaga, M.D. Network-Based Risk Assessment of the US Crude Pipeline Infrastructure, *International Journal of Critical Infrastructures*, 10, 1, 2014, pp. 67–80.

[2] (a) Hodges, L. The Quantum Physics of Emergency Management, *Journal of Business Continuity and Emergency Planning*, January 2019;(b) Hodges, L. Systems Fragility: The Sociology of Chaos, *Journal of Emergency Management*, May/June 2016;(c) Hodges, L. *Systems Fragility: The Sociology of Chaos*, Naval Postgraduate School, Center for Homeland Defense and Security, Monterey CA.

5

COMMUNICATIONS

The communications sector embraces all forms of electronic communications—communication satellite networks, landline telephone communications, radio, television, Geographical Position System (GPS) navigation, LORAN,[1] and wireless cellular and noncellular communications networks involving voice, data, and Internet. However, this chapter *does not* discuss broadcast communications such as radio, television, GPS, and LORAN. Rather, it focuses on the primary means of two-way, personal communications via telephones and the Internet.

This chapter describes how the communications sector is structured, how it works, its resiliency, and potential threat–asset pairs:

- *Three interconnected networks*: The three major telecommunications network infrastructure components are *landlines*, *wireless*, and *extraterrestrial* networks (communication satellites). These three provide a level of resilience through redundancy, but each one is vulnerable to physical, cyber, and high-powered microwave (HPM) attacks.
- *Multiple regulations*: Communications networks are primarily owned and operated by the private sector. These owners exert influence on the sector through the President's National Security Telecommunications Advisory Committee (NSTAC), which is a direct link to the executive branch of the US government. On the other hand, there is no clear-cut unity of government's role and responsibilities in the communications sector: it is regulated by at least three agencies: the Federal Communications Commission's (FCC) NRIC, Department of Commerce's National Telecommunications and Information Administration (NTIA), and Department of Homeland Security's (DHS) Cybersecurity and Infrastructure Security Agency (CISA).
- *1996 Telecommunications Act*: Communications has undergone radical restructuring since the enactment of the *1996 Telecommunications Act* that deregulated the sector. The primary result of this legislation has been the rise of the *carrier hotel*—the highly concentrated co-location of telecommunications and Internet equipment in one place. Co-location is a form of self-organization that restructured the infrastructure.
- *Carrier hotels*: Like the power and energy sectors, the communications sector is shaped by its transition from *vertical monopoly* to deregulated competitive *oligopoly*.[2] This regulatory reshaping produced network hubs (*carrier hotels* and large metropolitan exchanges), with high degree and betweenness centrality. Carrier hotels and their interconnection are the most critical assets in this sector.
- *Triple redundancy*: The three overlapping systems—landlines, cellular, and extraterrestrial—may be vulnerable to cascading cyber exploits because they are connected through a system of gateways. Thus a failure in one may lead to unexpected consequences in another. In general, redundancy may be a *disadvantage* in these

[1]LORAN is a network of land-based radio navigation beacons used by ships and aircraft to determine speed and position.

[2]A small group of controlling firms is considered an oligopoly.

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

highly connected networks if malware is allowed to cross from one to another.

- *Top 30 routes*: Critical nodes are clustered around major metropolitan areas linked by the top 30 landline routes connecting Chicago, Atlanta, Dallas–Fort Worth, San Francisco, New York, Washington–Baltimore, Los Angeles, Seattle, Denver, Sacramento, Philadelphia, Miami, Houston, Kansas City, Boston, Orlando, Portland, and San Diego. These 18 nodes and 30 links—plus connections to Asia, Canada, Europe, and South America—form a self-organized network with spectral radius of 4.73.
- *Submarine cables*: The major submarine cables circumventing the globe provide the backbone of global communications. Analysis of their betweenness criticality reveals the most critical links run through the Mediterranean Sea, Suez Canal, and Persian Gulf, out to the South China Sea to Singapore. A second most critical cable connects Florida, New York, and France. The top 10 most critical nodes in terms of betweenness are France, India, Florida, New York, Singapore, the United Kingdom, and Brazil, in descending order.
- *Risk and resilience*: The fundamental resilience line for the network containing the top 30 routes and carrier hotels places it in the high-risk category. The globe-spanning submarine cable network is also fragile due to high betweenness of cables linking Europe, the Middle East, Africa, India, and Singapore. In general, the communications sector is at extremely high risk of malware spreading because its spectral radius is very high. Both physical and cyber attack vectors raise the threat probability to nearly 100%.
- *Criticality*: More generally, the critical nodes of the communications sector are (a) carrier hotels, (b) IEC points of presence (POPS) and gateways, and (c) land earth stations (LES) that link communication satellites to terrestrial communications networks. LESs are particularly critical because there are so few of them. The critical links are high-bandwidth (OC-12 and OC-768) cables that link major carrier hotels together and the submarine cables connecting continents.
- *Unusual sector threats*: Besides cyber exploits, HPM and jamming threats pose unique sector-specific threat–asset pairs. Cyber–carrier hotel and cyber–exchange threat–asset pairs are asymmetric because of their low cost and capacity for virulent spreading. HPM–carrier hotel and cellular-jamming threat–asset pairs are also low cost and asymmetric, but often overlooked as viable threats.
- *Risk-informed strategy*: The optimal risk-informed strategy invests heavily in hubs (carrier hotels and exchanges) and betweeners (highly critical transmission cables). Optimal risk reduction and resiliency may be achieved by investing in the hardening or redundancy of hubs and large betweener links.

5.1 EARLY YEARS

The modern age of communications is rapidly transitioning from *analog* (information is encoded in a continuous signal) to *digital* (information is encoded as a stream of digits—zeros and ones). As a consequence, we think of analog as an old technology and digital as a new technology. But digital communications is far older than analog. The telegraph machine (1837–1873) was the first digital communications system because it coded information as a series of digits just as modern digital systems do. Western Union started the first transcontinental telegraph network in 1861 and introduced the hugely successful stock ticker in 1866. Because of its reach, Western Union and the railroads were responsible for establishing *time zones* across vast nations like the United States. Digital telegraphy was such a huge success that Western Union became the first communications monopoly by1866.

Telegraphy had one major drawback—it required a trained operator to translate the digital data into words and the reverse—words into digital code. This limited its usefulness as a consumer product. What people really wanted was a talking telegraph machine. (Western Union was forced out of the voice communications business in 1879 when it lost a patent lawsuit with Bell Telephone Company.)

Sound is analog. Sound waves travel through the air as a continuous wave form. Thus it seems only logical that a talking telegraph should encode sound (voice) as an analog signal—a continuous wave form. If only the energy of sound could be converted into electrical energy, transmitted as an electrical analog signal, and then converted back into analog sound at the other end, the telegraph could talk. Thus was born the idea of a telephone.

Alexander Graham Bell (1847–1922) demonstrated the first operating telephone in 1876. Bell combined his knowledge of speech therapy with contemporary theories of electricity to create the first voice telephony device to win a US patent. Bell was a contemporary of James Clerk Maxwell—the great Scottish scientist who formulated the rules governing electromagnetic fields. Both men were born in Scotland and educated in England. While Maxwell was a mathematical theoretician, Bell was a practical thinker. He liked to make things. His family moved to Boston in 1870 where he set up his speech and elocution school for training teachers of the deaf. He later became a professor of speech and vocal physiology at Boston University, specializing in teaching deaf-mutes to talk.

Another contemporary, Michael Faraday (1791–1867), demonstrated the principle of electromagnetic induction—the basis of converting electrical signals into audio by vibrating a membrane—and the reverse. So Bell combined these technologies into one: speech waves vibrate a membrane surrounded by an electric field, which induces a current in a wire. The wire transmits the oscillating signal to a far point where the process is reversed—oscillating current induces vibration in a membrane to create sound waves. Bell correctly reasoned that human speech could induce electrical oscillations in a wire and then be converted back into sound waves, if properly amplified along the way.

In January 1878 Bell demonstrated his invention to Queen Victoria while on his honeymoon in England. He promptly got an order to install a private line between Osborne House, on the Isle of Wight, and Buckingham Palace. By the end of 1878 there were 5600 telephones in the United States. By 1882, there were 109,000, and by his death in 1922, there were 14 million telephones in the United States. Bell's patents expired in 1894, but by 1897 he had moved on to the study of aeronautics.

Bell, his father-in-law Gardiner Hubbard, and Thomas Sanders formed Bell Telephone Company in 1877. They established their first telephone exchange in New Haven, CT (21 telephones and 8 lines), and began expanding it outward—initially to Chicago and eventually to San Francisco by 1915. Growth was rapid because Bell Telephone licensed its patents to others, thus attracting investments in local exchanges and "telephone companies." Soon, Bell Telephone Company became American Bell Telephone Company—and on its way to becoming a national enterprise.

Licensing revenue allowed American Bell to buy controlling interest in Western Electric in 1882. Western Electric put American Bell into the equipment manufacturing business. Licensing and equipment manufacturing soon led to network system building. American Telephone and Telegraph Company (AT&T) was incorporated as a subsidiary of American Bell Company in 1885 for the sole purpose of building long-distance networks. In 1899 AT&T reorganized as an IP holding company that would cycle through several iterations of the *competitive exclusion principle* between 1913 and the present.

From 1898 to 1924 the communications industry was engaged in a "communications war" because of competition and rapid technological change in the industry. For example, the automated exchange and self-dial telephone invented by funeral undertaker Almon Strowger made Bell's equipment obsolete. The *Strowger switch* eliminated the human switchboard operator. Strowger suspected a human operator was sending his funeral business elsewhere, so he eliminated the position!

Competition from upstarts fragmented the industry. By 1903 there were 2 million telephones from independent companies versus 1,278,000 from Bell. In addition, Bell Telephone had developed a reputation for high prices and poor service. AT&T fell on hard times as the Bell System faltered and bankers began to take over. J. P. Morgan gained control of the company and installed Theodore Vail as president in 1907. The rescue of the Bell System also marked the beginning of its downfall as an unregulated company—Morgan's monopolistic consolidation of the independents soon led to the regulation of AT&T by Congress.

Under Vail's leadership and Morgan's backing, AT&T became a vertically integrated monopoly by 1911. The Department of Justice (DOJ) sued AT&T in 1913, claiming it had violated the Sherman Anti-Trust Act of 1890. The lawsuit resulted in restricting, but not stopping, AT&T. The 1913 *Kingsbury Commitment*—an agreement between AT&T and DOJ—kept AT&T from buying independents without DOJ's permission, required AT&T to interoperate with independents, and forced AT&T to divest itself of Western Electric. But by 1924, AT&T owned 223 of the 234 independents!

Vail and Morgan believed in monopolies:

> For much of its history, AT&T and its Bell System functioned as a legally sanctioned, regulated monopoly. The fundamental principle, formulated by AT&T president Theodore Vail in 1907, was that the telephone by the nature of its technology would operate most efficiently as a monopoly providing universal service. Vail wrote in that year's AT&T Annual Report that government regulation, "provided it is independent, intelligent, considerate, thorough and just," was an appropriate and acceptable substitute for the competitive marketplace.
>
> The United States government accepted this principle, initially in a 1913 agreement known as the *Kingsbury Commitment*. As part of this agreement, AT&T agreed to connect non-competing independent telephone companies to its network and divest its controlling interest in Western Union telegraph. At several later points, as political philosophy evolved, federal administrations investigated the telephone monopoly in light of general antitrust law and alleged company abuses. One notable result was an anti-trust suit filed in 1949, which led in 1956 to a consent decree signed by AT&T and Department of Justice, and filed in court, whereby AT&T agreed to restrict its activities to the regulated business of the national telephone system and government work.[3]

Remnants of the early communications wars remain today. Local telephone companies—called local exchange carriers (LECs)—operated in restricted regions called local access and transport areas (LATAs) until 1996. Prior to re-regulation of the industry in 1996, it was illegal for LECs to cross LATAs without permission from the FCC. This hampered adoption of new technology because LECs were monopolies within their LATAs and there was only one long-distance company—AT&T.

[3]https://www.corp.att.com/history/

But vertically integrated monopolies have advantages too. Components worked across the country, service quality was high, and access was universally available. The *universal access* policy guaranteed telephone service to anyone at a low cost, because installation and maintenance costs were amortized across all users. Operating as a regulated monopoly, AT&T was able to serve 99% of the population regardless of where people lived. Rural as well as densely settled metropolitan areas received telephone service under the 1934 law. Universal service also brought a high level of standardization of handsets, switching equipment, and transmission lines. It was a period of relatively secure and resilient service.

AT&T was declared a *natural monopoly* from 1934 to 1996. But the company did not stand still for 32 years. Rather it went through a long period of divestiture. This long and winding road began with the *Communications Act of 1934*. Congress asserted its control over broadcast and telecommunication companies and established the FCC as regulator of airwaves and all things having to do with communications. It declared the electromagnetic spectrum public—not private—property. For example, all communications companies—including radio, TV, and phone companies—must obtain licenses from the FCC for broadcasting and to operate wired and wireless networks.[4]

In 1974 the DOJ began taking a long series of steps leading to divestiture and re-regulation of the natural monopoly set up by the 1934 law. A long-drawn-out lawsuit from 1974 to 1984 led to the breakup of AT&T in 1984. In a profound decision, the 22 wholly owned Bell operating companies were separated from AT&T. The resulting seven regional *Baby Bells* became competitive local exchange carriers (CLECs) and no longer operated as protected monopolies.

The Baby Bells were Nynex in New York and New England; Bell Atlantic, BellSouth, and Ameritech in the Midwest; and Southwestern Bell, US West, and Pacific Telesis in California and Nevada. Between 1984 and 1996, the competitive exclusion principle once again reigned as these companies went through acquisitions and mergers leading to only a handful of CLECs. As a result, the sector consolidated even further to a few carriers such as AT&T, Verizon, Sprint, and T-Mobile.

The next major step in divestiture came in 1996 with the *Communications Act of 1996*. This law replaced the 1934 law and introduced major changes to the infrastructure and its reliability. Its impact is still rippling through the industry today.

The 1996 law re-regulates the industry by forcing carriers to rent their networks to anyone wanting to start and run a telephone company. The idea is to open up long-distance transmission to local telephone companies and conversely to open up access to the "last mile" to long-distance companies. But it still limits ownership of cable TV, television, and radio stations to specific percentages in each region, and it sets pricing on some services. Most significantly, the law did not apply to cellular telephony, which became the dominant form of the "last mile" over the next decade.

Inter-exchange carriers (IECs) (the long-distance carriers) are now required to interoperate and share assets. AT&T can use the lines of Level 3, and Level 3 can use the lines of AT&T. This is called *peering* in the industry. Peering is also responsible for the creation of *carrier hotels*—multi-tenant facilities containing storage and switching equipment from competing telephone and Internet companies—which has accelerated self-organized criticality in the communications infrastructure. By co-locating computers, switches, and storage, communications companies can operate faster and cheaper, in addition to linking together their networks.

The 1996 legislation attempted to set prices at the local level as well as the long-distance level. But the CLECs won a court order that forced the FCC out of the local market. As a consequence, states can regulate prices—not the FCC. Today, the CLECs can establish peering charges except where competitors cannot agree. In case of disagreement, states have the right to set pricing. So, today the FCC sets the wholesale price of long-distance service, but the peering fee charged by the local carrier is allowed to float within the limits of state regulation. In practice, this means that both wholesale and retail prices are controlled.

Peering produced volatility in the industry not unlike the volatility caused by *wheeling* in the electric power sector. If carrier A rents an hour on carrier B's network, and then sells an hour to carrier B on its network, the net difference should be zero. But when MCI WorldCom reported peering charges as capital expenditures instead of expenses, and peering income as revenue, its CEO Bernie Ebbers was accused of falsifying the company's accounting. He was found guilty of fraud and sentenced to 25 years in jail in 2005.

Today, the old Bell System companies are called *CLECs*, and the long-distance companies—including the *Bell Long Lines* system—are called *IECs*. These players are linked together through a network of carrier hotels and transmission backbones. A hub-and-spoke architecture has emerged as preferential attachment increased self-organization. Additionally, we have a system that is shaped by years of cycling from regulated monopoly to deregulated oligopolies. Peering and competitive exclusion tend to concentrate assets—an emergent process that is also driving global communications toward self-organized criticality. This means the communications sector is extremely fragile with respect to malware as well as physical attack on critical nodes such as carrier hotels and backbone transmission lines.

[4]One notable exception: Wi-Fi networks operating under 1 watt are permitted without a license.

5.2 REGULATORY STRUCTURE

The first critical infrastructure legislation in US history was a by-product of the 1962 Cuban Missile Crisis, resulting in the creation of National Communications System (NCS). Negotiations between President Kennedy and Premier Khrushchev were threatened by "call completion" problems. It was not possible for the two leaders to simply pick up the telephone and place a call to anywhere in the world like it is today. In fact, Khrushchev was forced to use Radio Moscow to communicate, indirectly, with Kennedy, and Kennedy used a variety of means to circumvent the Kremlin bureaucracy:

> During this time, ineffective communications were hampering the efforts of the leaders to reach a compromise. Without the ability to share critical information with each other using fax, e-mail, or secure telephones such as we have today, Premier Khrushchev and President Kennedy negotiated through letters. Generally, Washington and Moscow cabled these letters via their embassies. As the crisis continued, hours passed between the time one world leader wrote a letter and the other received it. Tensions heightened. On October 27 and 28, when communications became urgent, Premier Khrushchev bypassed the standard communication channels and broadcast his letters over Radio Moscow.[5]

The so-called hotline established after the crisis was initially a Teletype set up in August 1963. Kennedy also established the NCS by executive order:

> Following the crisis, President Kennedy, acting on a National Security Council recommendation, signed a Presidential memorandum establishing the NCS. The new system's objective was "to provide necessary communications for the Federal Government under all conditions ranging from a normal situation to national emergencies and international crises, including nuclear attack". At its inception on August 21, 1963, the NCS was a planning forum composed of six Federal agencies.[4]

In 1978 presidential executive order EO 12046 consolidated two other communications agencies into the NTIA under the Department of Commerce. NTIA combined the White House Office of Communications Policy (OTP) with Commerce's Office of Communications. The principal role of NTIA was to sell spectrum to telephone, radio, and TV companies. But its involvement in communications has sometimes extended beyond marketing of airwaves. For example, in 1998–1999, NTIA played a major role in commercialization of the Internet.

A third big step in governmental oversight was taken in 1982 when President Reagan issued executive order EO 12382. This order established another watchdog organization that reported directly to the President—NSTAC. NSTAC members are senior management (CEOs and senior vice presidents) of telecom companies. Their job is to advise the President on matters of communications security.

PDD 63 (1998) designated the US Department of Commerce as the lead agency and the NTIA as the sector liaisons official for the information and communications sector. NCS was responsible for making sure the communications sector worked, the NTIA regulated the airwaves, and the NSTAC advised the President.

During the 1990s, the FCC became concerned with the *Y2K problem* (turnover of the millennium calendar that threatened to render computers and communications equipment inoperable). This prompted the FCC to temporarily create a National Reliability and Interoperability Council (NRIC) in 1993. It was dismantled after the Y2K threat subsided, but then the FCC rechartered NRIC in 2002, on the heels of 9/11. A series of reports issued by NRIC in 2002 remain an authoritative source of recommendations on how to secure the communications infrastructure. At this point, there were no less than four agencies overseeing communications: FCC/NRIC, Commerce/NTIA, DHS/NCS, and NSTAC.

NCS became part of DHS in 2003. It was tucked under the Critical Infrastructure Protection Division, which was eventually rolled up under the Cybersecurity and Infrastructure Agency in 2018. One of its objectives was to partner with the major communications owners and operators. For example, in 2003, these were AT&T, Cisco Systems, Computer Sciences Corporation, COMSAT Corporation, EDS, ITT Industries, National Communications Alliance, Nortel Networks, Science Applications International Corporation, Sprint, United States Communications Association, Verizon, and MCI WorldCom. Contrast this list with the members of NRIC: AT&T, Microsoft, Nokia, Nortel, Qwest, MCI WorldCom, Motorola, Alcatel, Sprint, Verizon, Lockheed Martin, Boeing, AOL-Time Warner, EarthLink, Level 3, BellSouth, DHS, NCS, Hughes, Intelsat, Communications Workers of America, Comcast, Cox Communications, Cingular, and Cable & Wireless.

President Obama consolidated these sprawling agencies (by executive order EO 13618) in July 2013 (see Fig. 5.1). At that time, the DHS oversaw the National Protection and Programs Directorate (NPPD), which oversaw the Office of Cyber security and Communications (CS&C), which oversaw the offices of Government Emergency Telecommunications Service (GETS), Telecommunications Service Priority (TSP), Wireless Priority Service (WPS), and Shared Resources High Frequency Radio Program (SHARES). This structure evolved further as threats evolved. By 2018, these sprawling agencies were consolidated under the CISA umbrella.

Emergency responders obtain authority and access to cellular networks through CISA and its subagencies.[6] The

[5]www.dhs.gov

[6]Standard Operating Procedure 303 (SOP 303) describes a shutdown and restoration process for use by commercial and private wireless networks in the event of a national crisis.

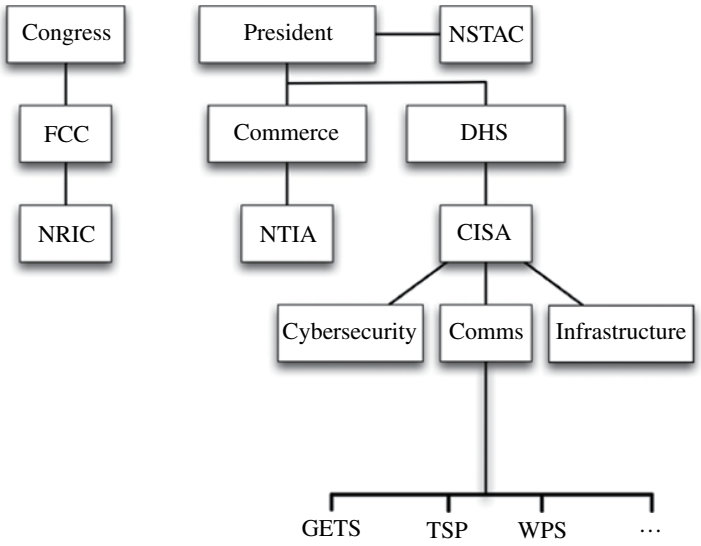


FIGURE 5.1 The structure of US governmental agencies involved in the regulation and protection of the communications sector.

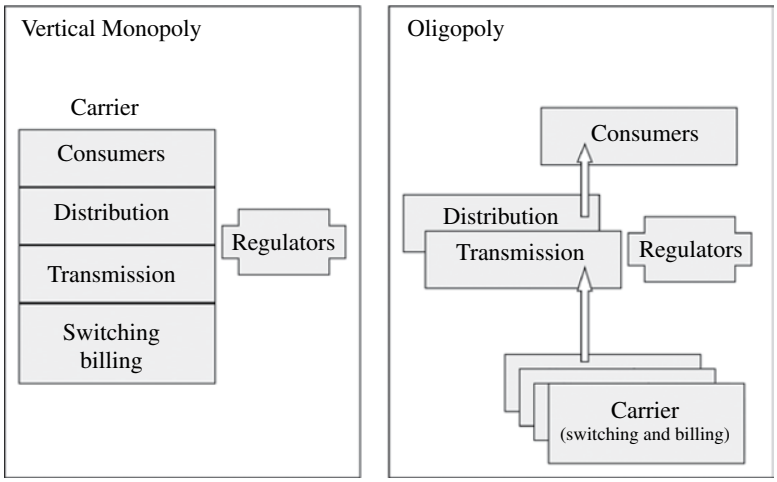


FIGURE 5.2 The 1996 Telecommunications Act deregulated the communications sector and is reshaping it into an oligopoly of competitive companies on a global scale.

GETS, TSP, and WPS offices issue priority pins to federal, state, local, and tribal governments to be used during an emergency or crisis situation when the various networks are congested and the probability of completing a normal call is reduced. When used, these pins remove congestion and give first responders the highest priority.

5.3 THE ARCHITECTURE OF THE COMMUNICATIONS SECTOR

The Communications Act of 1996 reshaped the communications sector by changing it from a *vertical monopoly* to re-regulated competitive *oligopoly* (see Fig. 5.2). Long-distance landline carrier service (IECs) is a price-regulated utility much like electric power, gas, and oil. However, unlike other sectors, there was a temporary surplus of capacity because of heavy investment in transmission lines during the dot-com bubble of 1995–2000. By 2012 however, this surplus was beginning to diminish, as new technologies replaced old.

The 1996 law created competitive local exchanges, required network peering, and placed caps on wholesale and retail pricing. Most profoundly, the previously proprietary transmission and distribution lines were cast into an industrial commons open to all competitors. The vertically integrated AT&T monopoly has been replaced by an oligopoly—many companies sharing the same switching and transmission commons that all telephony and Internet services depend on. Verizon, AT&T, T-Mobile, Vodafone, and others depend on the same infrastructure. But there is little incentive to maintain

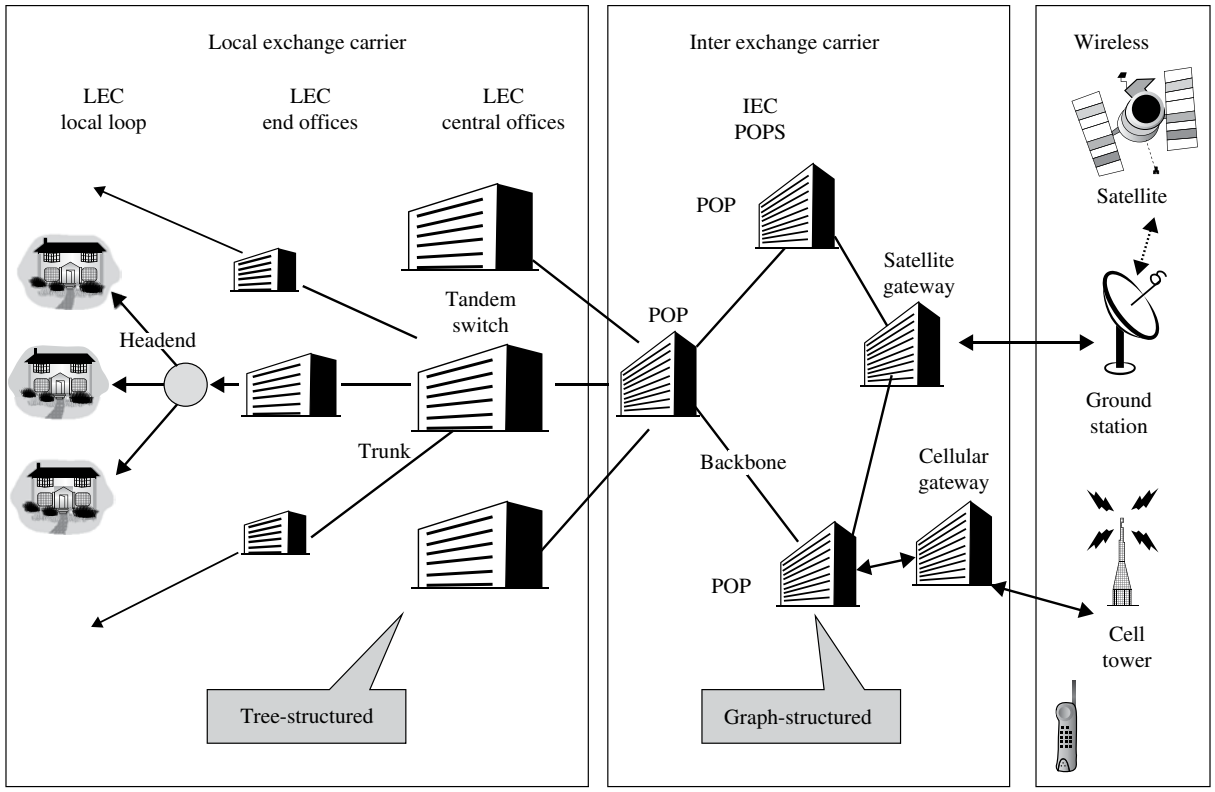


FIGURE 5.3 The architecture of the communications sector includes landlines, satellites, wireless, and access points.

this shared resource. Is this industrial commons sustainable or will rapid expansion impact its resilience?[7]

5.3.1 Physical Infrastructure

Against this background lies a huge infrastructure undergoing massive technological change. How does it work, and how do we derive security policies from these mechanisms? First, we must understand the basic terminology and architecture of the communications infrastructure.

Plain old telephone service (POTS) was capable of transmitting 64 kbps of digital data over copper wires. A POTS call requires 8 kbps of control, so computer users get 56 kbps of data when they dial up a POTS line and use a modem to connect their computer to the Internet. This standard has become the basic unit of bandwidth in the telephone network, designated DS0 (Digital Service Zero).

Circuits are combined to create more capacity. For example, a DS1, also known as a T1 line, is 24 DS0 lines working as one and yields 1.536 Mbps of data and 8 kbps of control. Therefore, a T1 line transmits 1.544 Mbps of data and control information. Similarly, a T3 (also known as a DS3) line is 28 T1 circuits plus control bits, yielding 44.736 Mbps overall.

[7]Government sets wholesale and retail prices, so what is the incentive to maintain and improve the long-distance transmission backbone?

Capacity goes up by combining circuits or changing technology. An optical fiber cable (OC) transmits more information than a copper cable. Optical transmission lines are designated as OC-1 (51 Mbps), OC-3 (155 Mbps), OC-12 (622 Mbps), and so on, up to OC-768 (Gbps). An ordinary cable TV coaxial cable can transmit from 3 to 10 Mbps. The very-high-bandwidth connections provided by OC-12 and OC-768 cables are of paramount importance to security and resilience of this sector because of their capacity to haul a lot of information. Flow resilience of these transmission lines will be one focus of risk analysis.

The global Internet is held together by an international submarine cable network spanning all continents (except Antarctica) with ultra-high-speed cable, typically in Tbps (terabits per second—or thousands of Gbps). Over 95% of all data transmission is carried by undersea optical cable. At the time this was written, there were 420 submarine cables stretching over 700,000 miles of ocean. The longest cable—SEA-ME-WE3—connecting Asia and Europe is 24,000 miles in length. It suffered service disruptions 16 times between 2005 and 2018. Most disruptions are due to breaks caused by fishing, ship anchors, and underwater landslides and volcanoes.

In the United States, copper and optical landlines are the backbone of the communications commons including a variety of technologies. They carry voice, data, Internet email, audio, video, and any other digital information that can be converted into short bursts called *packets*, tagged with

source and destination addresses, and routed through the infrastructure shown in Figure 5.3. Landlines are the fabric that holds it all together. They cross borders, tunnel beneath roads, oceans, railways, and buildings. Packet switching optimizes the use of transmission lines by sharing multiple "conversations" over one line.

Various transmission and switching technologies hold the communications sector together. Figure 5.3 is a simplification of the overall network architecture. It leaves out details such as microwave relay stations, laser links, and so forth. But it is sufficiently detailed for us to come to some conclusions about criticality and fragility. Generally, the sector is shaped by preferential attachment resulting from the 1996 Telecommunications Act.

The major functional purpose of the IECs is to provide long-distance connectivity and to connect the CLECs together into one national network. This is done by providing POPS, *network access points* (NAPS) for Internet users, and *gateways* for integrating satellite and cellular networks into the backbone, as shown in Figure 5.3. POPS, NAPS, gateways, and switching equipment typically reside in carrier hotels because of the peering requirement and cost efficiencies.

CLECs provide *local loop* service. They connect to the backbone through POPS that switch calls to their central offices. In turn, central offices funnel calls to end offices, which in turn channel the call to consumers through a neighborhood switch known as a *headend*. Headends handle approximately 1000 users at a time. CLEC networks are shaped like a hierarchical tree, while the IEC networks are arbitrary graphs—not necessarily a tree or grid. Note the redundancy in the local loop due to multiple tandem switches. The switching fabric at the CLEC level is relatively resilient and failures are localized to a few thousand customers.

5.3.2 Wireless Networks

Wireless transmission is governed by radio technology, which requires licensed spectrum for all but Wi-Fi. Various bands (frequency ranges or colors in the electromagnetic spectrum) have been set aside for cellular, satellite, and local area networks. For example, radios connecting Earth-orbiting satellites to ground stations operate at speeds comparable to POTS. The so-called 3G, 4G, LTE (Long Term Evolution), and 5G cellular wireless networks operate at megabit and gigabit speeds (see Table 5.1).

Wi-Fi is a special case, because it does not require a license to operate in most countries. It has a limited range due to limitations on its power. Wi-Fi networks linking together personal computers over short ranges are currently operating in the range of 100 Mbps but will exceed gigabits in the future. By 2020 5G deployment may supplant Wi-Fi, or the two may interoperate at gigabit speeds. Other technologies such as cellular and high-powered wireless can operate at much higher speeds or over longer distances. Each technology has its advantages and disadvantages, which is why they coexist in the marketplace.

TABLE 5.1 Evolution of cellular telephony

Generation	Era	Speed	Main features
1G	1980–1990	2.4 kbps	Analog voice, no security
2G	1991–2000	64 kbps	Analog voice, digital data, SMS, encryption
3G	2000–2010	2.0 Mbps	Analog voice, digital packet data, wireless access security
4G (LTE)	2010–2020	100 Mbps	Internet protocols, Wi-Fi, end-to-end security
5G	2020–?	1–2 Gbps	Interactive multimedia (video, telepresence, augmented and virtual reality)

Wireless transmission encompasses three major technologies: communication satellites, cellular, and Wi-Fi or WiMAX. Wi-Fi was formally called the IEEE 802.11 standard, but that designation was dropped in 2018. WiMAX was formally called the IEEE 806.16 standard deployed as 4G by some carriers. These various standards have subsequently been blurred by convergence to broadband access methods that utilize wider bands of spectrum to carry many bit streams in parallel. The underlying technology of 5G may render these other technologies obsolete.

Satellites have global coverage, but are relatively low speed and expensive. A number of low earth satellite networks (LEOs) deployed in the early 2020s may change the capability of satellite communications both technically and economically. These extraterrestrial networks operate as swarms surrounding the Earth at relatively low altitudes.

Cellular towers are relatively low cost, but lack complete coverage. The deployment of high-speed 5G with a range of one kilometer requires much denser small cells. No single technology serves all purposes. It is likely that by the mid-2020s the communications sector will be a mix of technologies integrated into a larger and more divers IEC backbone as shown in Figure 5.3.

5.3.3 Extraterrestrial Communication

There are thousands of communication satellites in use today and the number continues to increase by the thousands. While the public is mostly unaware of their presence in the national communications infrastructure, they play a critical role in voice and data communications, broadcast television transmission, military surveillance and imaging, intelligence gathering, early warning systems, maritime and aeronautic navigation with GPS, weather forecasting, inspection of agricultural lands, rescue and disaster relief, oceanographic and natural resource observations, and so on.

In 2004 satellite operators earned revenues of $2.3 billion—$1.4 billion for moving data and $900 million for providing voice services. By 2016 the satellite industry reported revenues of $260 billion, including launches. This is still small compared to the entire communications industry and is growing slowly compared to undersea cable. The communications sector as a whole generated revenues of $1400 billion. However, satellite communication remains very important to first responders and emergency management organizations because it provides wireless access from almost any place on Earth.[8]

The idea of communication satellites circling the Earth originated with science fiction writer Arthur C. Clarke in 1945. Clarke was way ahead of his time. His article described how a rocket circling the Earth at 22,300 miles above the equator would hover above the same land area, because it would circle the Earth at the same speed as the Earth rotates. The rocket is parked in a geosynchronous orbit. He recommended three geosynchronous rockets be stationed above the Earth at 120° apart so together they could cover all of the Earth's surface. Clarke invented the GEO (geosynchronous Earth orbit) satellite, which was actually constructed and put into orbit 20 years later.

Today there are three kinds of communication satellites: LEO (low Earth orbit), MEO (medium Earth orbit), and GEO (geostationary Earth orbit). GEO is the oldest, followed by MEO and LEO networks. Each has its advantages and disadvantages in terms of latency (time delay due to the time it takes a radio signal to make a roundtrip from Earth to satellite and back), bandwidth (the transmission speed), coverage (how much of the Earth's surface is served by one satellite), power (how much power it takes to send and receive the radio signal, and hence the size and weight of the handsets), and cost (how many satellites, how heavy, and how powerful). In simple terms, the further away a satellite is, the more surface it covers, but also the more power and larger size required to send and receive messages.

GEO satellites circle the Earth at 22,300 miles, which exactly matches the rotational speed of the Earth while simultaneously giving the satellite enough centripetal force to offset gravity. Hence they hover over the same location all the time, which also gives them large coverage. There were 402 GEO satellites in 2015, but this is considered "crowded" because of the necessary separation between geostationary objects. (Lawrence Roberts of the Berkeley Technology Law Review estimates the maximum capacity is 1800 geostationary satellites.[9])

Reservations in space are made by the *International Telecommunications Union* (ITU), which regulates GEO spectrum. In 1967, the *United Nations Outer Space Treaty* declared the geosynchronous orbit as a "common heritage of mankind." ITU determined that slots in this orbit were up for grabs on a first come first serve basis. A space rush ensued, and today the GEO orbit is considered a valued asset. In addition to limited slots, GEO satellites introduce a delay that complicates the transmission of Internet packets.

Inmarsat was the first and most successful GEO network system in the world. Started in 1979, its network consists of 13 satellites linked to the global telecommunications network through 34 LESs, all run from a network operations center in London, United Kingdom. Satellite coverage is 95% of the surface of the globe (north and south poles have no coverage). Inmarsat provides bandwidth comparable to 3G cellular to consumers with broadband access to the Internet.

Inmarsat service has been used to monitor radiation leakage in power plants and oil refinery monitoring in the energy sector. Asset tracking is a major application of satellite communications because of global coverage: GPS container tracking by shippers, equipment tracking by large farms, train and car tracking by railway operators, and vessel tracking of fleets at sea.

Satellites provide an alternate and redundant communications network. Because they work from outer space, they are available when landlines and cell phones are not. Hence they are especially important to emergency workers. For example, emergency satellite communication services (via Stratos, Inc.—a satellite service reseller) were employed after the 9/11 terrorist attacks on the Twin Towers:

> On Sept. 13, a Federal law enforcement agency contacted Stratos from the scene at Ground Zero in New York City, looking for a communications solution that didn't require land-based facilities. Stratos sent a shipment of Iridium phones to New York City, which arrived there hours after receiving the initial equipment request. After consulting with Federal officers at a command station a few blocks from the World Trade Center rubble in lower Manhattan, the Stratos team installed two Iridium fixed-site terminals on a nearby roof and another in a mobile command station. The equipment was used for emergency back up communications to help facilitate the agency's relief and damage containment efforts.[10]

5.3.4 Land Earth Stations

LESs handle bulk traffic between satellites and the terrestrial network. They are key assets in the IEC backbone because they handle large volumes of international phone calls, emails, and TV broadcasts. One of the oldest and largest LESs in the world is located at the southern tip of the British Isles—*Goonhilly Station*. It has 60 dishes spread across 140 acres in Cornwall. It transmits to every corner of the globe via space and through undersea fiber-optic cable:

> On 11 July 1962 this site transmitted the first live television signal across the Atlantic from Europe to the USA, via TELSTAR. This Satellite Earth Station was designed and

[8]Satellite communication does not reach the north and south poles.
[9]https://www.space.com/29222-geosynchronous-orbit.html
[10]http://www.stratosglobal.com/

> built by the British Post Office Engineering Department. Goonhilly-Downs covers 140 acres and is located at the westernmost end of the Cornwall coast in England. It was selected because of the topography of the land. The first satellite dish to be built on the site, Goonhilly-1, also known as Arthur, was an 85 feet in diameter parabolic design weighing 1118 tons. It set a world standard for the open parabolic design of the dish.[11]

Large LESs exist in the United States too. For example, the Staten Island Teleport, owned by Teleport Communications Group (TCG), handles much of the broadcast telecommunication streaming in and out of media capital Manhattan. The 100-acre business park includes a 400-mile regional fiber-optic network and an operations center linked to a satellite transmission facility.

5.3.5 Cellular Networks

The cellular telephone wireless network also feeds into the IEC backbone as shown in Figure 5.3. Cellular telephones have become a pervasive commodity—expected to exceed the population of humans on Earth some time before 2020. Their dependence on terrestrial landlines cannot be ignored, however. Without landlines connecting cellular towers to the architecture shown in Figure 5.3, cell phones are worthless. In addition, cellular service covers far less than 100% of land and sea area, because the range of a cell phone is approximately ½ (5G) to 3 (4G) miles.

Cell phones operate on only one standard in Europe (GSM), which means networks interoperate across country borders. But in the United States, cellular networks have grown up somewhat like the landline LECs grew up in the 1890s—as sprawling competitors. The result is an overly complicated and confusing cellular network infrastructure. In order to fully understand this important infrastructure, we have to delve into the arcane world of cellular access methods and technology generations—a topic beyond the scope of this book. However, these different methods of encoding and decoding signals in the airways are converging on 5G—ultimately a single global standard.

The cellular network derives its name from the fact that it is actually a honeycomb of regions called *cells*—each cell acting like its own self-contained radio broadcast network. These cells communicate with a tower located in the middle of the region. The tower links each handset to a wired network that interfaces with a gateway to the IEC backbone through a POP. Cells divide a city into small areas about 10 miles in diameter and automatically transfer communication links from one cell to another as the handset moves. Major highways and freeways are densely populated with towers to track consumers as they move through different cells.

As the world transitions to 5G, the size of cells will shrink down to ½ mile (1 km) radius and the number of small cells needed to connect with handheld devices will increase 10-fold. The deployment of 5G will take some time for this reason, but the benefits of low latency and high transmission speeds will promote rapid adoption of 5G. Table 5.1 shows how 5G completes the decades-long transition of cellular to Internet standards, essentially transitioning the Internet to a mostly wireless infrastructure.

Towers and their associated switching gear are called *base stations* and *small cells*. Each base station is connected to the Mobile Telephone Switching Office (MTSO), which ties into the wired phone system through a gateway (see Fig. 5.3). A base station tracks every handset as it moves in and out of cells. When the handset leaves one cell and enters another, the signal is handed off to the next tower. Switching is fast enough that users do not notice the gap as cell phones roam from one base station to the next without interruption.

A cell phone needs three numbers to operate within its cell, cross over into another cell, and interoperate with the wired landline network. Each phone has a System Identification Number (SID), a unique five-digit number assigned by the FCC to each carrier; an Electronic Serial Number (ESN), a unique 32-bit number programmed into the phone when manufactured; and a Mobile Identification Number (MIN), a 10-digit number derived from your phone's dial-up number. The SID validates that your phone is legal and works with the correct network, the ESN validates that you have registered with a carrier such as Verizon, and the MIN uniquely identifies the consumer.

Here is (roughly) how a cell phone works. When the handset is turned on, it is assigned one of 42 control channels to send its SID to the base station with the strongest signal. The MTSO switch monitors signal strength as you move from one cell to another. The connection is handed off to the cell with the strongest signal. If the handset SID does not match the SID of the base station, then the handset must be "roaming"—which means the caller is outside of his or her home base station cell. The MTSO that is handling the call uses the SID, ESN, and MIN numbers to track the handset and pass its signal on to a gateway into the IEC backbone and then to another MTSO. The receiving MTSO locates the destination handset through a reverse process and makes the connection. The switching equipment in each MTSO must be sophisticated enough to perform handoffs at both ends—without the consumer realizing what is happening.

5.3.6 Generations

The technological shifts that rapidly advance cell phone technology are known as *generations*—1G for the first generation, 2G for second generation, and 4G for the latest generation of

[11]https://en.wikipedia.org/wiki/Goonhilly_Satellite_Earth_Station#History

wireless phones (see Table 5.1). Similar to the rapid pace of personal computers in the 1980s and 1990s, early cellular technology eventually reached a level where consumers no longer care about the technology. However, the first five generations deserve mention here for historical reasons.

First-generation (1G) cellular networks ran on analog signals and are often called AMPS (Advanced Mobile Phone System). Second-generation (2G) phones converted sound into digital signals containing speech and data. So 2G cell phones introduced the first generation of *digital* telephony. An interim generation called 2.5G ran on an all-digital network but was capable of only supporting email, Web browsing, and low-resolution photos. The ITU defined 3G networks as wireless digital networks supporting transmissions from 144 kbps to 2 Mbps—roughly equivalent to wire line DSL. 3G also adopted packet switching protocols, which made it to conform closer to the Internet. Fourth generation (4G) and beyond combined bandwidth enhancements and features like movies on demand with apps, games, and eBook services and came even closer to true Internet.

5G advances wireless communications in several significant ways. First, latency—the time to connect and commences communicating—is fast enough to react in real time and play fast-moving games. Second, speed of transmission even under heavy congestion is 10–100× faster than 4G. This enables virtual and augmented reality and projection of holographic images in 3D. Third, 5G will eventually become the global and universal protocol for all wireless Internet. This has major implications for security because it creates a monoculture sameness: an attack on one 5G device is an attack on all devices.

Along the way to the Gs several other interim generations were deployed. For example, Cellular Digital Packet Data (CDPD) is digital data transmitted over AMPS networks. A few police departments may still use CDPD, because it is inexpensive and covers large and sparsely settled areas of the country. However, AMPS with CDPD is not very secure.

5.3.7 Wi-Fi Technology

Wi-Fi is the commercial name for a series of standards set by the IEEE 802.11x committee. Wi-Fi devices are small radios operating below one watt of power so they can operate on an unlicensed band. 802.11x is a series of technologies that progressed from 50 Mbps to over 150 Mbps. In 2018, the Wi-Fi Alliance stopped using the IEEE standards to enumerate generations of Wi-Fi. Instead, the industry alliance began using simple numbers. Wi-Fi 6 is renamed IEEE 802.11ax.

Wi-Fi may eventually be replaced by 5G, but keep in mind that Wi-Fi runs on an unlicensed band, while 5G requires FCC approval. This may have the effect of making Wi-Fi an indoor wireless infrastructure and 5G an outdoor infrastructure. Cell phones operate over both technologies.

Wi-Fi's first encryption algorithm, the Wired Equivalent Privacy (WEP), was easy to break, so the Wi-Fi Protected Access (WPA, WPA2) algorithms created in 2003 and 2004 quickly eclipsed the original standard. Only computers, cell phones, and tablets set up using the 2006 Wireless Protected Setup (WPS) standard are guaranteed to be secure, although some implementations are not secure as of 2019. The WEP protocol subsequently went through a number of enhancements to address vulnerabilities. WEP2, WEPplus, and dynamic WEP are some of the enhanced versions. Open-source tools like Nessus and Nikto may be used to scan for vulnerabilities in Wi-Fi networks.[12]

Wi-Fi has a very short range (100 m), but relatively high speed and low power requirements. In 1985 the FCC allowed Wi-Fi broadcasts without a license, which means access points can be installed anywhere, by anyone. Additionally, the technology was simple enough to be produced at low cost. This propelled Wi-Fi to mainstream use, not only in offices but also in restaurants, libraries, shopping malls, and other public places. Almost all computer, tablet, and cell phone devices contain Wi-Fi chips.

5.4 RISK AND RESILIENCE ANALYSIS

The redundancy provided by the three major telecommunications network infrastructure components—landlines, cellular, and extraterrestrial networks (communication satellites)—add resilience to the sector because service can be switched from one to the other during an emergency. Landline and cellular service can be backed up by satellite communication services, for example, and Wi-Fi can sometimes complete the last mile connection when landlines fail. Wi-Fi riding on tethered balloons reestablish access to the Internet following disasters. During emergencies, Wi-Fi access points running at 5 W or more can achieve ranges of 20–30 miles.

The optical fiber infrastructure is relatively robust due to redundant paths, but much of the long-distance wiring still depends on the old AT&T long lines or its silhouettes. These pathways were designed to be efficient and not necessarily redundant. Moreover, the 1996 Telecommunications Act has driven Internet and telephony topology to a state of self-organized criticality today. Peering—and self-organizing *preferential attachment*—has produced highly concentrated *carrier hotels* at strategic locations around the country (see Fig. 5.4).

Generally, threat–asset pairs in the system of Figure 5.4 are the following (in order of criticality):

1. Terrorism, power, cyber attacks on telecom hotels.
2. Terrorism, accidental damage to submarine cables.
3. Terrorism, power, cyber attacks on LESs.
4. Weather, terrorism, power, cyber attacks on IEC POPS

[12]https://en.wikipedia.org/wiki/Cracking_of_wireless_networks#Nessus

FIGURE 5.4 Major carrier hotels within the United States form the backbone of telephony and Internet service.

5. Terrorism, power, cyber attacks on cellular gateway POPS.
6. Power outages on CLEC central offices.
7. Terrorism attacks on satellites, towers, cables, and fiber.

Human-caused hazards are likely to be from:

- Cyber attack on all telecomm components—terrestrial and extraterrestrial.
- Severing of undersea cables.
- Physical attack on carrier hotel—destruction of concentrated assets.
- Physical attack on LES—damage to a critical link.
- HPM attack on telecomm components.
- Physical attack on IEC POPS; gateways.
- Physical or HPM attack on satellite "bird."

Natural causes of hazards are likely to be from:

- Weather
- Power outages
- Component failure

Figure 5.5 shows a sample fault tree risk model of the sector. Assuming individual threat and vulnerability probabilities are both 50%, the entire sector is 92.5% likely to fail due to one or more threat–asset pair hazards. Indeed, telephony outages prior to the 1996 legislation were long-tailed hazards with a risk profile as shown in Figure 5.6. Kuhn studied outages for two years from April 1992 to March 1994 and showed that the pre-deregulation era telephone system was 99.99% reliable [1]. The top three hazards responsible for the most downtime (%) were:

1. Overloaded circuits (44%)
2. Human error (28%)
3. Acts of nature—weather (18%)

The top three sources of failures by cause (%) were:

1. Human error (49%)
2. Hardware failure (19%)
3. Software failure (14%)

Figure 5.6 presents Kuhn's data in the form of exceedence probability and risk profile versus consequence measured in millions of customer-minutes—the length of time communication outages left consumers without service. True exceedence is calculated as described in Appendix B. Short outages are much more likely than long service drops. This is shown in both exceedence probability and risk profile. Risk starts out high and steadily declines versus customer-minutes. Fractal dimension is 1.5, placing these normal accidents solidly in the low-risk category.

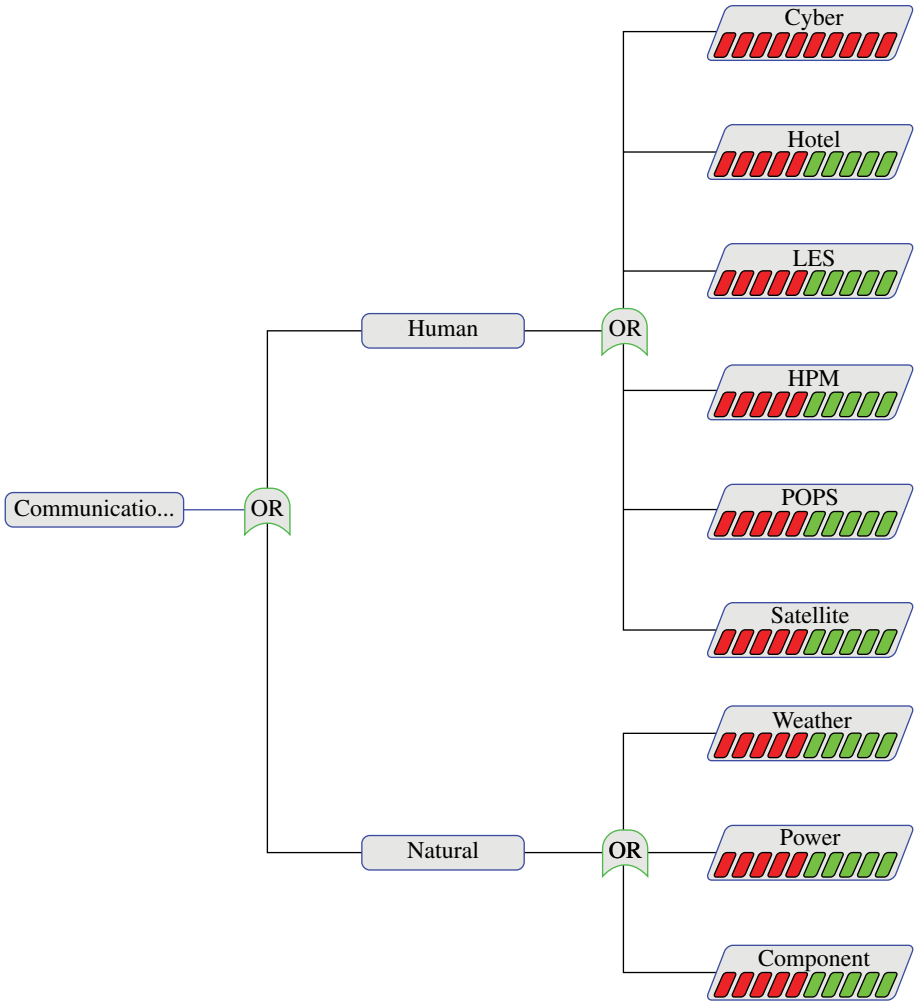


FIGURE 5.5 Human-caused hazard fault tree risk model for the communications sector lists the most likely threats.

The risk profile of Figure 5.6 is unusual because it starts relatively high and steadily declines. Most risk is small—at the low end of the consequence axis—and high-risk outages are very rare. The fact that the risk profile approaches zero as consequence increases without bound confirms the low-risk hypothesis. Prior to the 1996 Telecommunications Act, the POTS was comparatively resilient. Kuhn attributes this mainly to loose coupling of the network. The communications network contained little of what Perrow called *catastrophic potential* and Bak called *self-organization*. But that all ended with the emergence of carrier hotels.

5.4.1 Importance of Carrier Hotels

For economic reasons as well as the Telecommunications Act of 1996, CLECs and IECs, communications companies, Internet ISPs, and businesses were motivated to co-locate equipment and services in the same building. This saves money, because infrastructure costs can be amortized over a large number of tenants. They are attractive to carriers because they provide:

- High-speed connections (fiber, satellite, microwave)
- Roof access for antennas
- Physical security
  - Key card access
  - Video surveillance
  - Biometric scanners
- Power and backup generators
- VESDA air sampling (imminent fire detection)
- Fire suppression—suppressors and sprinklers
- Redundant HVAC
- Seismic strength

Because these functions are expensive and bothersome for businesses to supply on their own, many carrier hotels also contain key assets outsourced by their clients. Cloud computers, databases, and so on are often co-located in a carrier hotel. Carrier hotels also provide wireless gateways, storage and hosting servers for businesses, and application service

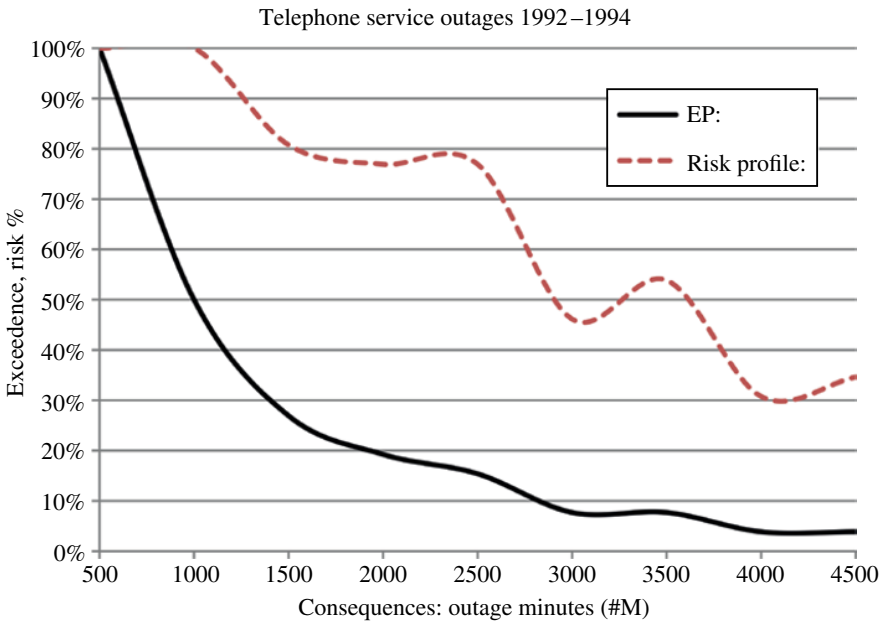


FIGURE 5.6 Telephone outages reported by Kuhn indicate the US communications infrastructure in the mid-1990s was low risk.

providers (companies that run your applications for you). If a carrier hotel is vulnerable, then the businesses that co-locate in them are vulnerable as well.

The largest carrier hotels in the United States—60 Hudson Street in New York and 1 Wilshire Boulevard in Los Angeles—became carrier hotels in large part because they happened to sit on top of a big optical fiber intersection. Like gigantic Internet onramps, these carrier hotels provide rapid access to far points of the globe. One building, Number 1 Wilshire Boulevard, is home to nearly 100 telecommunication carriers alone and is sometimes described as a *direct jack* to Asia and Japan. The large building at 60 Hudson Street in Manhattan houses switching equipment that connects the United States to Europe, Middle East, and Africa. Similarly, the Weston Building in Seattle connects Canada and Alaska to the lower 48, and the carrier hotel in Miami links the United States to South America.

Richard Clarke—the first cybersecurity head at the DHS—recognized the importance of carrier hotels early on:

> I'm told … that although Transatlantic Fiber lands at about 10 different places in Massachusetts, Rhode Island, Long Island and New Jersey that, after having landed, it all goes to one of two facilities—60 Hudson Street or 111 Eighth Avenue in Lower Manhattan. If that's true, that would seem to be a problem. … I suspect this statement … is true, that if you blew up 60 Hudson Street and 111 Eighth Avenue, we could not communicate via fiber optic with Europe.[13]

Carrier hotel criticality was one of the immediate concerns expressed by President Bush on the heels of 9/11. NSTAC's report to the President avoided an alarmist call to action, but it identified carrier hotels as critical components of the communications sector:

> Although no analyses performed to date have shown that the entire communications architecture would be adversely affected through the loss of a single telecom facility, according to JPO-STC, loss of specific communications nodes can cause disruption to national missions under certain circumstances. As a result of these analyses, the JPO-STC not only has shown the dependencies of Department of Defense (DoD) missions on communications, but also reports that there are further and more far-reaching implications to other national infrastructure sectors. [2]

5.4.2 NETWORK ANALYSIS

Carrier hotels are the most important assets because most of the communications infrastructure depends on them—wired as well as wireless. For example, the IEC POPS and gateways—for both satellite and cellular—are typically co-located in a carrier hotel. They tie the entire sector together and handle most of the traffic. This is why a risk-informed strategy focuses on high-capacity transmission links and highly concentrated carrier hotels—especially carrier hotels with high influence or connectivity.

Consider the busiest routes circa 2003 in the top-level communication infrastructure in the United States as shown in Figure 5.7. This network connects the major carrier hotels in the United States. Chicago is the hub of this network and also the node with the largest betweenness value. Its connectivity is 9 and its 207 paths give it the highest betweenness rating. Betweenness and connectivity combine to yield a normalized weight of 4.0—making it the central node of this

[13]Richard Clarke, March 11, 2002.

network. In rank order according to normalized betweenness and normalized weight, the critical nodes are:

Chicago (4.0)
Atlanta (3.19)
Washington–Baltimore (2.69)
Dallas–Ft. Worth (2.47)
New York (2.44)

The links (routes) with the highest betweenness (normalized weight) are:

Atlanta–Washington–Baltimore (1.60)
Seattle–Chicago (1.53)
Chicago–New York (1.51)
Atlanta–Miami (1.51)
Atlanta–Chicago (1.49)

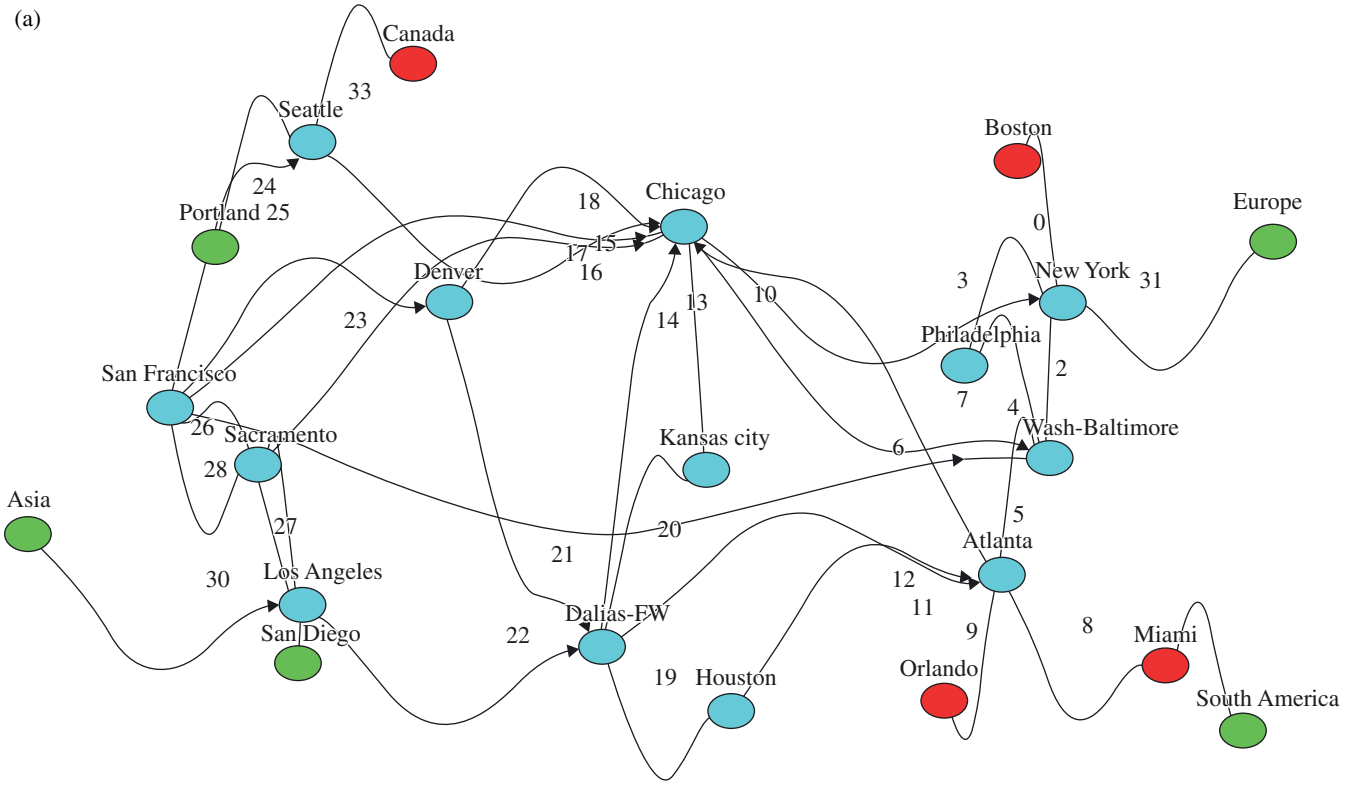


FIGURE 5.7 Critical factor analysis and resilience of the top 30 telecom routes in the United States circa 2003 using hypothetical parameters. (a) Top 30 routes and carrier hotel network for the United States circa 2003 has a high physical connectivity and betweenness node centered in Chicago. (b) Cascade resilience (6.24) of the top 30 routes assuming a node vulnerability of 25%. (c) Flow resilience (1.45) of the top 30 routes assuming a directional network. Reversing the direction of links yields an even lower flow resilience of 0.8.

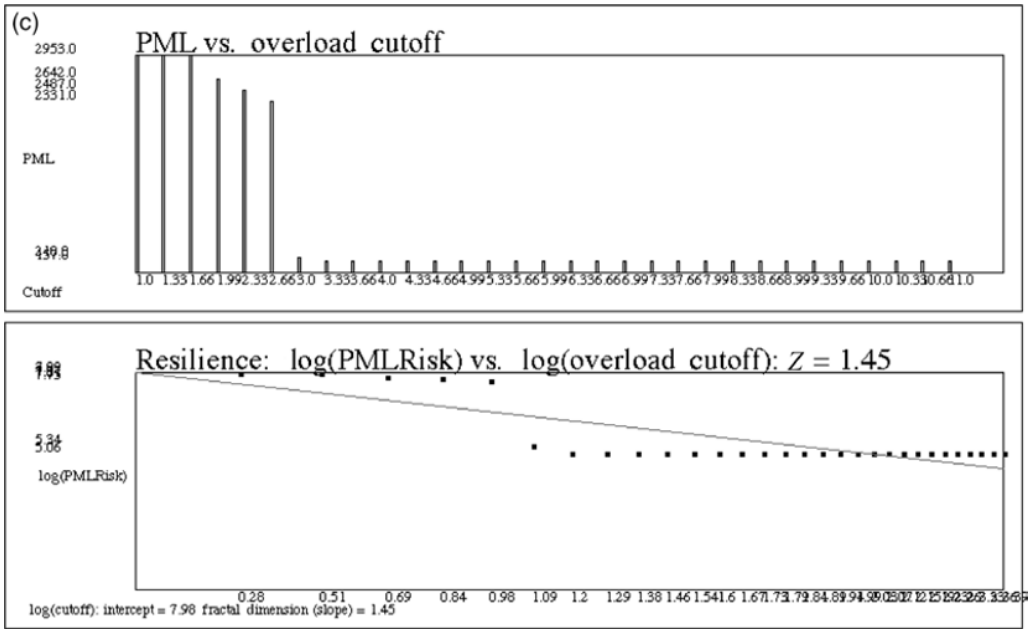


FIGURE 5.7 (Continued)

Cascade collapses might be caused by the spread of a computer virus or by backed up congestion due to a carrier hotel failure. Alternatively, a failure in one node may propagate to adjacent nodes with similar results. Simulated random attacks on hubs and targeted attacks on the hub (Chicago) were tallied to obtain the fractal dimensions and fundamental risk and resilience lines for each (see Fig. 5.7). The spectral radius of this network is 4.73, suggesting a fair amount of self-organized criticality. Cascade resilience, assuming node vulnerability to cascading of 25%, is good—6.24 on a scale of 0–10. But flow resilience is low, ranging from 0.80 to 1.45, depending on the direction of flow.

Targeted attacks on the hub are three times more effective than random attacks in causing cascade failures. But if the hub at Chicago is protected (hardened), risk drops and resilience increases. Random attacks on the network—with a protected hub—yields 50% more cascade resilience than random attacks and four times better than targeted attacks on the hub.

Similarly, PML risk is reduced by 40%, assuming consequence is the same across all nodes. Protecting only the hub shortens the long-tailed exceedence probability distribution and reduces risk of cascade failure for all nodes—not just the hub. The entire network is made more resilient by hardening only one of the 18 domestic nodes. The return on investment is greatest when protecting critical nodes.

5.4.3 Flow Analysis

Figure 5.7c shows how rapidly PML risk rises as the cutoff value of overload decreases along the *x*-axis and approaches 1.0. This suggests there are not enough alternate paths or re-routing causes overloading, which in turn causes one or more links to fail. The network contains four blocking nodes and five blocking links, which suggests there is potential for bottlenecks and high levels of betweenness.

Betweenness analysis assuming one link has failed so that re-routing is required yields overload ratios greater than 1.0 on 25 of the 30 links. The bottlenecks range from overloading of 1.17 to 8.66, suggesting extreme fragility when one link fails. The route from Sacramento to San Francisco is the most critical and the link between Sacramento and Los Angeles is second in order of betweenness criticality. San Francisco, Sacramento, Washington–Baltimore, Dallas–Fort Worth, and Chicago—in descending order—are the most critical nodes relative to potential bottlenecks (Link 1 betweenness ratios range from 1.86 to 1.37).

This example illustrates the paradox of redundancy. Adding a link between Portland and Sacramento increases spectral radius from 4.70 to 4.73, with a corresponding decrease in cascade resilience—6.24–6.19. PML risk of cascading increases from 330.0 to 350.9, with a corresponding decrease in number of blocking nodes (from 4 to 3), and lowers the maximum bottleneck ratio from 8.66 to 7.66. The paradox is that cascade resilience declines with an added link, but flow resilience improves.

5.4.4 Robustness

Link robustness is 43% (40% by approximation formula), which means that 43% of the links can be removed before the network is separated into islands. Since there are 30 links, 13 can be damaged or dropped without separation. Therefore, 17 links are critical. Obviously, all links connecting nodes with a single link are critical. There are

five blocking links that are essential to keeping the network together:

Los Angeles → San Diego
Atlanta → Orlando
Atlanta → Miami
Seattle → Portland
New York → Boston

Similarly, node robustness is 77% (79% by approximation formula), which means removal of any one of 14 nodes (0.77(18)) will not separate the network. Alternatively, four blocking nodes are critical because removal of any one of them will disconnect the network. They are Seattle, Atlanta, Los Angeles, and New York. Removal of these four nodes separates this CIKR network into six components, namely, Boston, Portland, Miami, Orlando, San Diego, and everything else:

- The top betweenness nodes appear in the center: France, Egypt, Djibouti, and India.
- The top 10 nodes in terms of critical factor betweenness normalized to the interval [0, 1].

5.4.5 The Submarine Cable Network

Figure 5.8a shows a 168-node network model of the 251 major submarine communication cables spanning the globe. By far, most voice, data, and Internet traffic travels through this network compared to extraterrestrial communication. These cables form the backbone of global communication.

The network is somewhat self-organized with a spectral radius of 5.45—182% of the average connectivity but with high cascade resilience of 7.47, assuming node vulnerability of 25%. The critical nodes and links lie along a path of high-betweenness nodes and links shown in Figure 5.8b. France and Egypt are the betweener nodes of keen interest because of their high bottleneck betweenness.

The most critical paths start with Egypt and extend forward and backward as follows. Normalized bottleneck betweenness is shown in parentheses:

Egypt (1.0) → (0.715) Djibouti → (0.497) India → (0.303) Singapore
Egypt (1.0) → (0.937) France → (0.349) New York → (0.547) Florida

This result concurs with an independent result reported by John Crain using a different technique. In his thesis, Crain concludes, "the passage through Egypt from the Mediterranean Sea to the Red Sea is a natural choke point. The same is true with undersea cables" [3]. Bottleneck betweenness is an easy but labor-intensive method of assessing risk and resilience in a network. A network of this size, however, requires an automated software tool.

5.4.6 HPM Attacks

One unusual and important threat to communications merits further analysis here. HPM guns are low-cost energy weapons that cause havoc when unleashed on computer, telecommunication, radar, and other electronic devices. A burst of energy from an HPM gun "fries" the circuits of most electronic machines. While carrier hotels are well protected against physical attack, they may be susceptible to cyber and HPM attacks.

HPM waves are created by discharging extremely short bursts of microwaves at high energy levels—typically gigawatts of energy fired in nanosecond bursts. These waves are short, that is, from a few meters to a few centimeters in length or from 100 MHz to 10's of GHz in the frequency domain. This is the electronic equivalent of a sharp knife, cutting through walls and shielding to get to electronic circuits. These attacks would damage machines but go unnoticed by humans.

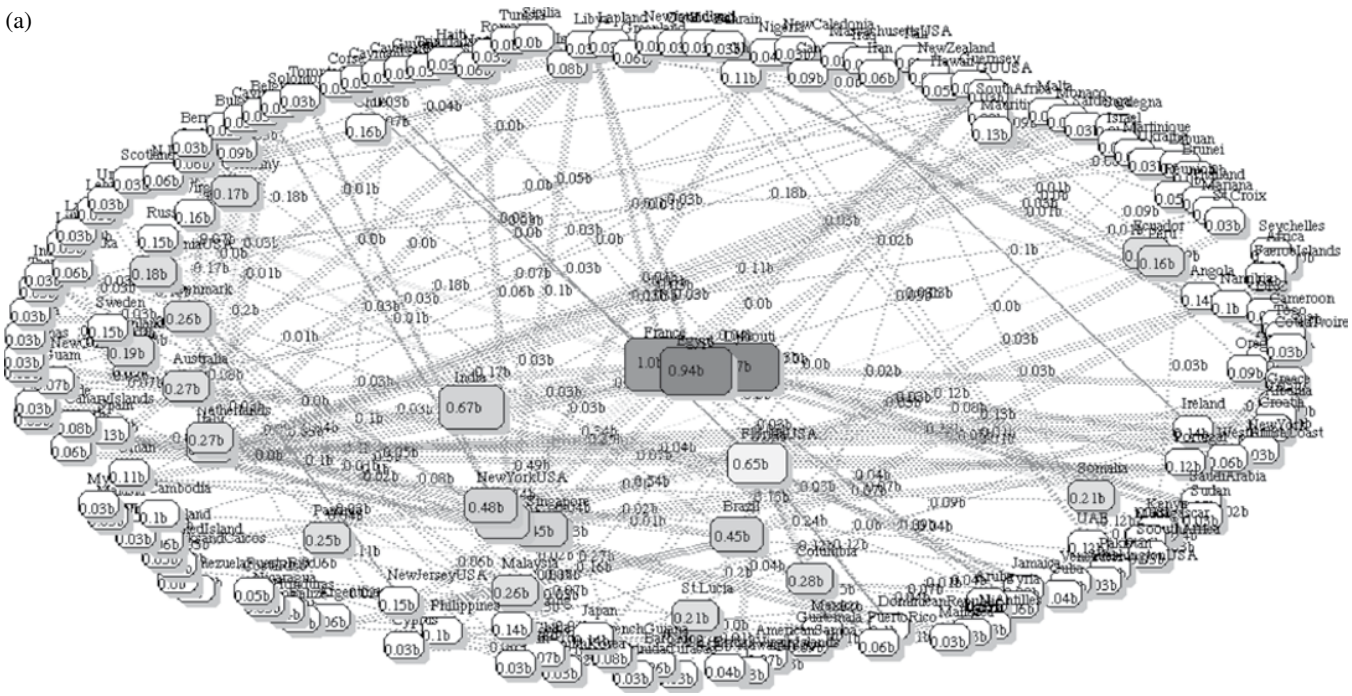
One way to think of HPM is to make an analogy with a high-heeled shoe. If the area of the heel is one square inch and the person wearing the shoe weighs 100 lb, the heel presses against the floor with a force of 100 lb/in^2. Now, if the area of the heel is reduced to one-half of a square inch, the 100 lb pressure is spread across one-half as much area, so the downward pressure is 200 lb/in^2. If we continue to reduce the size of the heel, say, to one-tenth of a square inch, the force against the floor is now 1000 lb/in^2! If we apply the same weight to a smaller and smaller area, the force goes up and up. A person that weighs 100 lb can apply a million pounds per square inch by simply wearing extremely pointed high heels. This is the idea behind HPM—energy is discharged over a very brief time interval, producing a large force—for a brief moment. But the force does not have to last very long to render damage.

HPM devices are made from a variety of components—all of which can be purchased from almost any electronics store. In addition, there are a variety of methods for storing and discharging "work" in extremely short bursts, ranging from magnetic to electronic linear accelerometers.

HPM weapons are ideal for asymmetric attacks on computer and electronic equipment, because they are:

- Silent and easy to conceal.
- Easily transported by truck, van, or even briefcase.
- Difficult to locate and destroy.
- Effective against nearly any unshielded electronic device.
  - Penetrate many materials.
  - Damage may not be apparent.
  - Not necessarily harmful to people.

(a)



(b) Submarine cable bottleneck criticality vs. node

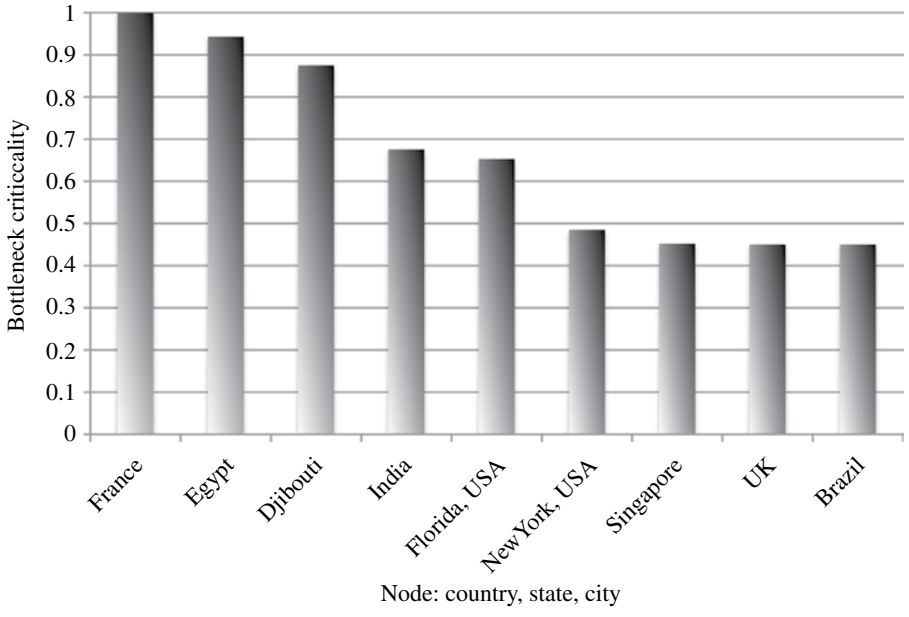


FIGURE 5.8 The major submarine cables circling the globe carry most of the communication traffic moving from country to county. (a) The network model centered around betweenness shows France as the most critical. (b) The top 10 nodes in terms of betweenness include the most critical links and nodes.

HPM attacks are asymmetric—they can do a lot of damage, but cost very little to build and deploy. They are portable, and most people cannot identify them. Weapons like HPM can penetrate the best physical defenses of most carrier hotels and do more physical damage than cyber exploits.

5.5 CELLULAR NETWORK THREATS

As cellular telephones become more and more powerful miniature computers, and the cellular network becomes more and more like a wireless Internet connection, the threats become more and more like cybersecurity and

Internet threats. This is the dark side of convergence—the use of Internet protocol (IP) in all communications sectors including TV, radio, and cellular telephony. Threats that work against IP networks spread to all converged networks including factory control networks, banks, and transportation. Disruption of an automobile from a cell phone has been demonstrated and is a real threat to drivers. Convergence to 5G will contribute even more criticality due to the enormous monoculture it creates.

The major threats to cellular networks fall into three categories: cyber, HPM-like, and physical. The top contenders are:

1. Cyber threats
   - Denial of service—flooding the airwaves with messages.
   - Disruption of control—taking over a control system.
   - Cloning—intercepting the phone's SID and MIN.
2. HPM-like threats
   - Radio-frequency (RF) jamming—blocking out the signal.
   - Damaging electronic equipment
3. Physical threats
   - Destruction of base stations—bombing.
   - Gateways and POPS—bombing of carrier hotels.

5.5.1 Cyber Threats

Security experts call cyber assaults *exploits*. One of the most common exploits is known as *denial of service* (DoS) or distributed DoS (DDoS) because it renders the network useless by overloading the channel with meaningless messages. A *DDoS attack* is like sending millions of automobiles onto a freeway to prevent ambulances and police cars from using the roadway. DDoS in cellular networks works the same way—by overwhelming the network with calls, thus rendering the network useless during an emergency.

DDoS attacks are not theoretical—they actually happen all the time. And because cell phones are integrated with computer networks, a DDoS attack can spill over into other parts of the communications infrastructure, and vice versa. For example, Spain's Telefonica cellular network was attacked by "SMS bombing"—a short message system DDoS attack—in June 2000. Flooding of the Spanish cellular network was actually a side effect of an email virus—called *Timofonica*—that spread through computer networks, infected address books, and then dialed cell phone numbers at random. Timofonica contaminated copies of *Microsoft Outlook* using a *macro* that randomly generated and dialed the phone numbers:

> Timofonica was marketed as a cell phone virus when in actual fact it was simply a clever variant of the good old email virus. Victims received an email with an exploitative attachment. When the attachment was executed an email was sent to every entry in the victim's address book and an SMS message was sent to random cell phones on the Telefonica network in Spain. The SMS message did not erase any critical information from the phone or cause any damage to the phone's operating system. It didn't spread from phone to phone. It was merely a variant of the spam we receive every day in our email inbox. [4]

Cloning—stealing phone identities and using them on unregistered handsets—is a far more insidious cyber attack. Analog cell phone identities are snatched out of the air, as crooks use small electronic radio scanners to intercept cell phone transmissions. Later, they use the encoded information to "clone" a second phone, billing their calls to the account of the phone that was scanned. This exploit diminished as analog phones gave way to digital handsets. But it is still a viable threat to police and emergency personnel that still use analog communications.

SS7 Skylock (Fig. 5.3) showing the overall architecture of the communications sector contains cellular gateway switches for bridging the gap between cell towers and the IEC network. One particular gateway is designated SS7 (Switching System 7) responsible for connecting cell phone messages to the communication backbone. The SS7 tracks where every cell phone is (its nearest tower or access point) and facilitates call completion at both ends.

Skylock is software for tracking cell phones and managing SS7 waypoints. It knows a lot about cell phone identification numbers, carrier numbers, and caller's location and phone number. Accordingly, it is a target of malicious hackers, especially criminals looking to defeat two-factor authentication (2FA). For example, most secure online systems use two factors to authenticate users. A password and cell phone number for the two factors. During login, an e-commerce site will send a pin number to a consumer's cell phone. The consumer must enter both password and pin number to complete the login.

Passwords and pin numbers must go through the SS7 switch to connect consumer with an online service such as a bank. In addition to the location and identity of the consumer, the SS7 temporarily knows the phone number used to complete 2FA. The SS7 provides an opportunity for man-in-the-middle attacks, when the SS7 switch is hacked.

A criminal can intercept the cellular call containing the consumer's pin number via the SS7 switch. Together with the password and pin number, the criminal can access bank accounts, credit card accounts, and any online account that relies on 2FA. Skylock is certified software used by the communications industry, but when compromised by hackers, it becomes a potent hacker tool.

Many more cyber exploits exist in the wild—even weaponized viruses like *Stuxnet*—which will be explored in greater detail in the following chapters. In general, these

exploits are getting more sophisticated as authorities and *black-hats* engage in an arms race to outdo one another.

5.5.2 HPM-Like Threats

HPM-like *RF jamming* is the process of blocking wireless transmission by sending out an interfering signal that cancels the true signal. These illegal devices can be easily purchased for less than $1000 from companies around the world: a few examples are Special Electronic Security Products, U.K. Ltd. of Manchester, England; Intelligence Support Group, Ltd. based in China Lake, Calif.; and an Israeli company called NetLine, manufacturer of the C-Guard. A portable C-Guard sells for about $900. Another company offers the $890 M2 Jammer, which comes in a briefcase and can block phones within a radius of 50 ft. Hub-Giant of Taipei, Taiwan, sells its WAC1000 personal jammer, which has an operating radius of up to 30 ft, for $169. And Uptron of Lucknow, India, offers a full range of jammers with coverage ranging from 20 ft to over 1 mile.

Manufacturers of jammers claim they are selling their devices to give anti-cell phone advocates a little peace and quiet from the ring of cell phones—especially in public places: "Cell phone jammers are readily available on the Internet. Many can be battery-powered and fit in a pocket or briefcase for people who would like to enjoy a meal, movie or church service in peace" [5].

5.5.3 Physical Threats

Physical threats are the least sophisticated and yet bombs are by far the most preferred weapon of terrorists. Gateways and POPS are typically concentrated in carrier hotels, so these become obvious bombing targets. But other physical threats—even more asymmetric—may be employed. For example, chemical attacks against major carrier hotels are not out of the question. Similar threats confront large and unprotected LESs, such as those located at Staten Island, NY, and Niles Canyon, CA.

5.6 ANALYSIS

Redundant tandem switches and ring structures in local loops as well as some IEC loops provide a degree of security due to redundancy. In addition, the abundance of long line fiber across the country suggests that there is sufficient redundancy in the backbone. But the top 30 routes were shown to be vulnerable to disruption of network continuity simply because they carry such a large proportion of all traffic. Accordingly, we have argued for protecting the most active metropolitan hubs, because assets are concentrated there and this is where traffic levels are the highest. They are high-value targets.

Asymmetric energy weapons such as HPM guns and RF jammers already exist and are proliferating. It would be a mistake to discount the threat of attack on communication and computer infrastructure from these weapons, simply because we know little about them. They already exist and are relatively inexpensive to acquire, hard to trace, and potentially very damaging.

The increasingly more concentrated hubs called carrier hotels exacerbates the communications sector's vulnerability to physical attacks. Since the Telecommunications Act of 1996, the physical resiliency of the Internet has slowly eroded because of bad policy and cost–benefit economics. The 1996 Act requires Internet Service Providers (ISPs) to peer—share each other's networks. While intended to encourage competition, the unintended consequence is a dramatic increase in vulnerability. In fact, NSTAC declared peering carrier hotels the number one physical vulnerability of the communications infrastructure:

> The current environment, characterized by the consolidation, concentration, and collocation of telecommunications assets, is the result of regulatory obligations, business imperatives, and technology changes, …Loss of specific telecommunications nodes can cause disruption to national missions under certain circumstances." [6, pp. 3]

Economics plays a major role in weakening the sector. As communications becomes increasingly commercialized and regulated, it also becomes increasingly efficient and cost-effective. Companies like AT&T, Level 3, Verizon, Hurricane Electric, Comcast, and Time Warner eliminate redundancy in the name of efficiency and profitability. What is wrong with efficiency? Efficiency translates into lower redundancy. Lower redundancy means more risk. And more risk means more vulnerability to hacks, outages, and exploits.

For the past two decades, carrier hotels have been putting more eggs into fewer baskets—fewer centralized facilities and fewer cables to carry the global load. Even if they do not reduce the number of paths available in physical cabling, hidden blocking nodes are very difficult to discover without massive computational resources. But this is only one issue. It is more cost-effective to increase the bandwidth (or number of cables) along an established path than to add alternative paths. Furthermore, connectivity of highly connected nodes is increasing, while connectivity of less connected systems stays the same or decreases. In other words, high bandwidth and high connectivity attracts more use, which attracts more bandwidth and connectivity, which attracts more use. This self-organizing feedback loop leads to fewer but larger-capacity nodes and connections. It reduces network robustness and resilience.

The bottom line is the physical Internet is evolving away from resiliency toward fragility. Think of it as the Interstate

Highway System where a major freeway is removed every few months to cut costs and optimize traffic by routing more and more traffic through the most traveled cities and freeway links. Even if more lanes are added to the overloaded roads, the fact that there are fewer roads leads to bigger catastrophic failures when one is blocked or damaged.

Communication Sector Strategy*: Self-organization in the form of preferential attachment has resulted in the accumulation of communications sector assets in a handful of carrier hotels, metropolitan area exchanges, and high-betweenness cables. The optimal risk-informed strategy invests heavily in these hubs and betweeners. In addition, the threat includes highly asymmetric HPM weapons as well as cyber exploits and physical attacks.*

5.7 EXERCISES

1. The most critical nodes in the communications sector are:
   a. Headends
   b. Critical fiber routes
   c. Network transport elements
   d. Carrier hotels/multi-tenant facilities
   e. Data centers
2. Which one of the following is an emergency telecom service?
   a. NSTAC
   b. GETS
   c. RSVP
   d. Communications sector-specific agency
   e. NRIA
3. Cellular and other wireless networks depend on which of the following for call completion?
   a. Wired landlines
   b. Wi-Fi access points
   c. Towers with a range of 100 miles
   d. Carrier hotels
   e. Satellite ground stations
4. Which of the following was responsible for deregulation (some say re-regulation) of the communications sector?
   a. The breakup of AT&T in 1984
   b. The Telecommunications Act of 1935
   c. The Telecommunications Act of 1996
   d. The Tragedy of the Commons of 2003
   e. The creation of NCS in 1963
5. Which of the following is legally responsible for the cybersecurity of the communications industry?
   a. NTIA
   b. NSTAC
   c. NCS
   d. NCC
   e. None of the above
6. Which of the following is the most critical component of communications, from a network resiliency point of view?
   a. Headends
   b. Local loop service
   c. IEC/POPS network
   d. CLEC network
   e. Wi-Fi access points
7. Communication satellites orbiting the earth were first envisioned by:
   a. Hedy Lamarr
   b. Alexander Graham Bell
   c. Theodore Vail
   d. President John F. Kennedy
   e. Arthur C. Clarke
8. There are currently three kinds of satellites in operation today: which of the following describes these kinds of satellites?
   a. Wi-Fi, 802.11
   b. LES, Goonhilly, and Staten Island Teleport
   c. LEO, MEO, and GEO
   d. Inmarsat, Marisat, and Westar
   e. Telstar, Intelsat, and Satcom
9. The International Communications Union declared orbits in space as:
   a. The common heritage of mankind
   b. The final frontier
   c. The property of the United Nations
   d. The property of property of Inmarsat
   e. There can be no more than GEO satellites
10. The largest LES in the world is:
    a. Goonhilly Station
    b. Staten Island Teleport
    c. NASA–Huston
    d. NASA–Cape Kennedy
    e. Arthur, named after Arthur C. Clarke
11. One of the earliest telephone installations was:
    a. Osborne House
    b. Niagara Falls
    c. White House
    d. Pentagon
    e. Seattle to Chicago
12. One of the earliest attempts to regulate AT&T was:
    a. The Telecommunications Act of 1934
    b. The Telecommunications Act of 1996
    c. The Kingsbury Commitment
    d. The Baby Bells
    e. The 1984 Accord
13. The electromagnetic spectrum used by cellular and Wi-Fi networks is:
    a. Public property
    b. Private property
    c. Intellectual property

d. Personal property
e. None of the above

14. The 1996 Telecommunications Act requires:
    a. Peering
    b. Wheeling
    c. Carrier hotels
    d. Digital data
    e. WiMAX
15. Which one of the following is considered a little known or understood threat to the communications sector?
    a. Floods
    b. Earthquakes
    c. Regulation
    d. HPM
    e. Tragedy of the commons

5.8 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. Is the US communications sector a cascading or flow network? What does resilience mean in terms of the spread of malware? What does resilience mean in terms of physical damage to a carrier hotel?

B. As the Internet becomes more wireless and as wireless transitions to 5G, does risk and resilience increase or decrease? Why?

C. What is the impact of increasing bandwidth along a communication path? Does it increase or decrease risk?

D. The 1996 Telecommunications Act requires peering. What is the impact of peering on self-organization and ultimately risk and resilience?

E. Are submarine cables increasing at a more rapid rate than satellite communications? Why or why not?

REFERENCES

[1] Kuhn, R. Sources of Failure in the Public Switched Telephone Network, *IEEE Computer*, 30, 4, April 1997, pp. 31–36.

[2] The President's National Security Telecommunications Advisory Committee. Vulnerabilities Task Force Report Concentration of Assets: Telecom Hotels, February 12, 2003, pp. 3.

[3] Crain, J.K. Assessing Resilience in the Global Undersea Cable Infrastructure. *Naval Postgraduate School*, June 2012, pp. 54. Available at https://calhoun.nps.edu/bitstream/handle/10945/7327/12Jun_Crain.pdf?sequence=1&isAllowed=y. Accessed July 21, 2019.

[4] McDonough, C. *Identifying the Risk Involved in Allowing Wireless, Portable Devices into Your Company*, SANS Institute, 2003, pp. 6.

[5] Wylie, M. Cell Phone Jammers, Illegal in U.S., Can Create Silent Zones, 2000. Available at http://www.newhouse.com/archive/story1a092200.html. Accessed June 27, 2014.

[6] NSTAC. Task Force on Concentration of Assets: Telecom Hotels, National Security Telecommunications Advisory Committee, February 12, 2003.

6

INTERNET

The *Internet* uses the communications sector to link computers, cellular telephones, tablets, transportation systems, water and power systems, and industrial control systems together. It extends the communications sector to applications such as the World Wide Web (WWW), email, video streaming, and smartphone apps like face-to-face conferencing and photograph-sharing services. A more precise technical definition says the Internet is a global network that uses Transmission Control Protocol/Internet Protocol (*TCP/IP*). Any device that connects to other devices via TCP/IP becomes part of the Internet by definition.

While the Internet has been around for over 50 years, it began to spread like an epidemic after TCP/IP was created and adopted by the US Department of Advanced Research Projects Agency (ARPA) (and later Defense Advanced Research Projects Agency [DARPA]), and the US government declared it open and available for commercialization in 1998. TCP/IP is the lingua franca of global communications and the basic "building block of the Internet's DNA." Open access and free open source software (FOSS) accelerated the explosive adoption of the Internet within a few decades beyond 1998.

The Internet started as an idea on paper and grew into one of the largest man-made machines in the world. Experts fully expect that all 7.5 billion inhabitants of the globe will eventually be on the Internet—perhaps by the mid-2020s. But the spread of TCP/IP goes beyond the population of the planet. Billions more machines communicate via the Internet from automobiles, factories, power grids, gas and oil pipelines, water systems, and everyday products. This proliferation of connected devices at the edge of the Internet has become known as the Internet of Things (IoT). The Internet will eventually connect more machines together than people—perhaps as many as 100 billion machines and 7.5 billion people will be linked via TCP/IP in the near future.

The global spread of TCP/IP and the confluence of computing and communicating are such an enormous topic that we devote an entire chapter to it. First, we briefly review the history of the Internet to prepare for following chapters on network and computer security. One of the most significant aspects of the Internet is the way it came into existence and the culture that supports it. Curiously, the Internet has no centralized governing body, although this is changing. It is an open community of globally distributed users that govern themselves. In some countries it is falling under governmental control. For example, large swaths of the Internet are banned in China, and the European Union (EU) enacted strong security and privacy legislation in 2018 called General Data Protection Regulation (GDPR). The GDPR is a set of regulations for protecting user's data and requiring the storage of user's data to co-locate in the same country as users. The future of the Internet is regulation.

Nonetheless, the sociology of this self-organized community is as interesting as the technology itself. The early highly decentralized volunteer developers and promoters of the Internet are being replaced by formal governing bodies. In the early days, a Request for Comment (RFC) continuing a technical innovation was circulated to volunteer experts who eventually came together to vote on the RFC. If approved and adopted by the majority of developers, an RFC

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

would be elevated to Best Common Practice (BCP) and immortalized as such.

The ad hoc volunteer system began to break down following the commercialization in 1998. Governance went global, and while the original volunteer organizations continued to operate, official governmental organizations began to infiltrate governance. In democratic societies, this took on the form of regulation and oversight. Authoritarian governments became more controlling, and they learned how to use the power and persuasiveness of the Internet to spread "fake news" and misinformation farther and faster than any other time in history. The impact that social networks have had on public policy, loss of privacy, and spread of propaganda is described in more detail in Chapter 9.

It is necessary to understand the basic communication principles of the Internet before embarking on the subject of cybersecurity, privacy, and regulation. However, if the reader has already mastered these basics, he or she may skip this chapter.

This chapter covers the following topics:

- *Internet Age*: Even though the Internet is much older than the personal computer (PC), it was not commercialized until 1992–1998. After that, it coevolved with the adoption of the consumer PC. Without the PC, the Internet may not have spread as explosively and globally as it is today, and conversely, without the Internet, the PC and smartphones might not have become as ubiquitous as they have. This coevolution propelled global societies into an Internet Age characterized by extremely high connectivity, short time intervals between events, adaptability and flexibility, and the rapid spread of global epidemics of ideas, political and social movements, and propaganda, as well as movement of products through the globalized world.
- *Non-secure*: PCs and the Internet are inherently vulnerable: the hardware and software of online computers are the first link in information technology security. A breach of software security in one computer can spread, like an epidemic, to millions of other information systems—all through the global connectivity provided by the Internet. While the Internet was designed to be redundant, it was not designed to be secure. And neither were the gateways to the Internet—the household PC, cell phones, tablets, industrial control systems, transportation systems, and other infrastructure that depends on the Internet. The rise of machine-to-machine connectivity, called IoT, vastly exacerbates the problem by increasing connectivity, spectral radius, influence, and vulnerabilities to massive cascading of malware.
- *TCP/IP defines the Internet*: The Internet is equivalent to the TCP/IP standard: networks that communicate in TCP/IP are considered "the Internet," and conversely, the Internet is considered as any network that requires the use of TCP/IP. TCP/IP is rapidly becoming the universal protocol for electronic communication. Unfortunately, TCP/IP is notoriously non-secure. It was not designed for secure communication. Rather, it was designed to be open and free, with an emphasis on resilience under nuclear attack. Cyber exploits and cyber war were never envisioned when the early Internet was created.
- *The Internet is not new*: The Internet grew out of ARPANet, which was a product of the Cold War: in 1969 the ARPA began a project that created the first "Internet" called ARPANet; ARPANet begat NSFNet and then merged back into NSFNet. The National Science Foundation (NSF), which ran the NSFNet for a time, was directed by the US Congress to commercialize the NSFNet in 1992. "The Internet" has become the consumer name of commercial NSFNet. The Internet grew at an explosive rate following its release by the US government in 1998.
- *Packet switching*: The biggest idea regarding the Internet is that data should be packet switched rather than circuit switched as it had been for over 100 years in the telephone network. Packets are blocks of data that contain their destination and return addresses so they can travel through the Internet on their own. Packet switching is much more flexible and efficient than circuit switching. Packet switching eventually became the default protocol for all kinds of communication including wireless cell phones, IoT, Wi-Fi, and 3G cellular. (Satellite radio and TV remain the only major communication signals that use streaming rather than use a packet protocol.)
- *Pioneers*: The Internet was invented by many people: Lickliter (the visionary); Taylor (the manager); Baran, Davies, and Kleinrock (packet switching); Postel (names and addresses of users); Cerf and Kahn (TCP); Tomlinson and Roberts (email); Crocker (governance); Metcalfe (Ethernet IP); Postel, Mockapetris and Partridge (DNS); and Berners-Lee (WWW). This handful of pioneers also governed the early Internet through a small collection of volunteer organizations.
- *RFC process*: Up until 1998 most of the decisions regarding operation and technology adoption were governed by a handful of volunteers. Decisions were made through the three-step RFC process. Step 1 required a written proposal that is circulated throughout the community. Step 2 requires a vote, and step 3 elevates a widely adopted RFC to BCP status. The Internet is self-documented—search on RFC 1234 or BCP 1234 and any user will obtain a link to the document.
- *Redundant but not secure*: The Internet was designed to be redundant: there are many alternate routes; packets are retransmitted when an error occurs; the global network rebuilds itself every day by updating a tree-structured

network of DNS servers; TCP is a protocol that automatically routes packets around broken lines and reorders packets when necessary; and the Internet has its own built-in SCADA system called SNMP for monitoring the devices on the Internet. However, the Internet was *not* designed to be secure. Version 6, called IPv6, encrypts packets and forms the basis of virtual private networks (VPN) running over public Internet. IPv6 became the standard in July 2017.

- *Graphical browsers made it popular*: The killer applications that ignited explosive growth of the Internet are email and the WWW. Marc Andreessen and Eric Bina created the first graphical user interface WWW browser in 1993, called MOSAIC, which set off consumer demand for WWW products and services throughout the world. Subsequently, many browsers have been developed and freely given to consumers, each with different levels of security and pop-up blockers, but browsers remain one of the principal vectors for the spread of malware.
- *Digital convergence*: The Internet has unified and standardized the coding of all forms of digital information. The User Datagram Protocol (UDP) is used instead of TCP/IP for streaming video, and encryption is an add-on to both TCP/IP and UDP. Email follows the higher-order rules of SMTP (Simple Mail Transport Protocol) and documents disseminated by the WWW follow the rules of HTML (Hypertext Markup Language), http (Hypertext Transport Protocol), and XML (eXtensible Markup Language)—universal standards for the encoding and transmission of text, pictures, sound, motion pictures, and animations. Unfortunately, these standards also form a monoculture—an exploit that succeeds on one device is an exploit that succeeds on all devices.
- *An unregulated infrastructure*: The Internet is not owned by anyone, and most governments have been reluctant to regulate it because of economic benefits that accrue from the Internet. Originally operated and governed by its users—corporations and volunteers who exerted influence through an open process called the RFC—it is now facing increasingly restrictive regulation from governments around the globe. Most decisions regarding Internet policies and standards were vetted through the Internet Society (ISOC) and its affiliated working groups such as the Internet Engineering Task Force (IETF) and the World Wide Web Consortium (W3C). Subsequently, the UN and authoritarian governments have restricted access and implemented various forms of regulations on users and Internet companies.
- *The Internet's backbone has hubs*: The WWW is vulnerable to attacks on its hubs—primarily through the so-called Tier-1 Internet Service Providers (ISPs), root servers, top-level domain servers (gTLDs), and highly connected e-commerce servers. These are collectively called *autonomous systems* (AS) and, like the Interstate Freeway system, carry the bulk of Internet traffic using the BGP (Border Gateway Protocol). Almost all of the traffic passes through these AS, which makes them attractive targets for malware. Therefore, even though there are billions of nodes in the global Internet, fewer than several hundred AS matter as far as the spread of malware is concerned.
- *The AS network is percolated*: A network analysis of the top 500, 1000, and 2000 AS reveal a very high level of self-organization—the Internet's spectral radius is high, and extremely high connectivity hubs are especially critical. As the number of AS-level ISPs considered rises from 500 to 2000, the spectral radius also rises, suggesting a high level of percolation. This means the Internet's AS network has a very low tolerance for the spread of malicious software. The Internet is fragile due to percolation and a handful of extremely highly connected super-spreaders.
- *A monoculture or fragmented Internet?* Globally accepted standards have emerged across the Internet, which forms a monoculture. At the same time, the Internet is threatened by balkanization—splintering of access and protocols behind a Chinese wall, an Iranian wall, a North Korean wall, or a paywall set up by e-commerce sites that require subscriptions and passwords. This raises the specter of fragmenting the Internet. Regulations like the GDPR may lead to further fragmentation as different regulations in different regions force e-commerce sites to implement different policies and apply different security and privacy technologies.
- *Rise of wireless*: Over the next decade or two, wireless access will become the default technology for Internet access. In particular, the 5G protocol will further unify and standardize communication as described in Chapter 5. The rise of 5G may see the fall of non-Internet technology such as cable TV, satellite radio and TV, and Wi-Fi. This will further exacerbate the vulnerability of a monoculture. One protocol used everywhere means malware is free to spread everywhere.

6.1 THE INTERNET MONOCULTURE

It should be noted that the Internet is the sum total of all TCP/IP networks that connect to one another. The WWW, on the other hand, is an application that runs on the Internet. The WWW is to the Internet what Microsoft Office is to Microsoft Windows—merely an application running on top of the Internet protocols (IP) described here. It should be noted that the WWW is not alone—other applications such as iTunes, Instagram, and Netflix are also applications

running on top of the Internet. So, the Internet is a type of operating system underneath many of the Web applications and smartphone apps in common use by billions of people.

These separate Internet ecosystems coexist in a symbiotic relationship, but they are not to be confused, one for the other. Some flaws leading to loss of security are inherently woven into the Internet fabric of protocols and software, while other flaws are woven into apps and applications running on top of the Internet. This chapter focuses on the Internet infrastructure and not the entire ecosystem of apps and applications. These will be described in Chapters 7 and 8.

The Internet is viral—it spreads to every part of modern life because TCP/IP is open and free and barriers to entry are nearly zero. In most countries, the cost of Internet registration is less than the price of a meal, and yet the Internet provides instant access to the world. But it has its downside: the Internet is embarrassingly open and non-secure. It is alarmingly a *monoculture*, and it is highly *percolated*. Disruption works both ways—bad people can use it to disrupt the lives of good people as easily as good people can use it to improve people's lives. Unfortunately, the Internet was designed not only to be easy to use but also easy to hack.

The Internet was a by-product of the *Cold War* between the former USSR and the West. The US government created the ARPA in response to the launch of Sputnik in 1957. The "missile gap" helped elect John F. Kennedy to the Presidency, and soon afterward, the US launched the Space Program that put the first men on the moon. But there was one smaller step taken on the journey to the moon in 1969 that may have been just as important. What was to become the Internet was "invented" by employees of ARPA who funded academics that built the first experimental Internet called ARPANet.

ARPA was, and still is, created for taking giant leaps forward to keep the United States technically ahead of its opponents. ARPA later become DARPA and initiated other forward-thinking ideas that would alter the world, but in its earliest days, it was focused on how to beat the Russians into space. The United States needed advanced computing capabilities—among other things—to make space exploration happen. The public relations similarity between the formation of ARPA and the formation in 2003 of the Department of Homeland Security is undeniable:

> All eyes were on ARPA when it opened its doors with a $520 million appropriation and a $2 billion budget plan. It was given direction over all US space programs and all advanced strategic missile research.[1]

In 1962, J. C. R. Lickliter moved from MIT to head the command and control program at ARPA. Lickliter surrounded himself with colleagues from Stanford University, MIT, UC–Berkeley, and UCLA—whom he dubbed the "Intergalactic Computer Network" group. In a memo to the Intergalactic Computer Network group 6 months after his arrival, Lickliter expressed frustration with the lack of interoperability and standards among computer centers:

> Consider the situation in which several centers are netted together, each center being highly individualistic and having its own special language and its own special way of doing things… is it not desirable or even necessary for all of the centers to agree upon some language, or at least, upon some conventions for asking questions as "What language do you speak?"[2]

Thus was born the idea of networked computers. But it would be Lickliter's successor, Robert Taylor, who took the next important step. Taylor was frustrated with having to login to three different computers from three different computer terminals—the so-called terminal problem. Instead of using separate terminals for different computers, why not link all computers together through a network and access each one from a single terminal? Computers should be just as easy to access as it is to call home through the telephone network.

Taylor convinced his ARPA boss to fund his project, arguing that his project would save money by solving the "terminal problem." A nationwide university network would make it possible for researchers all over the country to share expensive mainframe computers. In 1965, computers cost millions of dollars—a price barrier that prevented many academics from using them. But if a few expensive mainframes were made accessible via a network, then thousands of researchers could share the limited number of expensive machines. In 1968, ARPA contracted Bolt Beranek and Newman (BBN) to build ARPANet—the first version of what would become the Internet.

Meanwhile, others were thinking similar thoughts. One of the most profound ideas occurred to two people at about the same time. Paul Baran and an Englishman named Donald Davies both came up with the concept of a *packet*—"message blocks" of data that could travel through a network on their own rather than be harnessed to a single circuit. Telephone networks were *circuit switched*, which meant that they communicated by connecting the sender and receiver together via a dedicated electronic circuit. The entire circuit was consumed for the entire conversation. And only one pair of users could use the circuit-switched connection at a time. This is very inefficient.

Instead, a *packet-switched network* can share its wires or radio waves with packets from many users—all at the same time. Packets find their own way through a network and are

[1]Hafner, Katie, and Mathew Lyon, "Where Wizards Stay Up Late: The Origins of the Internet," Simon & Schuster, 1230 Avenue of the Americas, NY, NY, 10020, (1996), 304 pp. ISBN 0-684-81201-0. p. 20.

[2]Ibid, p. 38.

extremely efficient and flexible as compared to circuits, because multiple packets—all going to different destinations—can share a single circuit. This form of multiplexing made existing wires thousands of times more efficient.

Packets are "smart," because they contain their own source and destination addresses, much like a letter that is sent through the US Postal System. At each branch in the network, routing tables provide directions for where each packet should go next. Even if a portion of the physical network fails, an alternate path can be found and the packet rerouted. A simple algorithm was employed, called Open Shortest Path First (OSPF), which worked exactly as its name implies—by sending packets along the shortest, least congested paths. In this way, data communication becomes robust—a failure in one part of the network cannot disable the entire network.

While at UCLA working on an ARPA contract, Leonard Kleinrock proved packet switching to be superior to circuit switching. His theoretical analysis reinforced the intuition of Baran and Davies. Not only was packet switching a good idea, but it was now theoretically sound. The stage was set for a revolution in data communications. But change takes time, so the obscure ARPANet would take a few more decades to realize its potential.

By 1969 the ARPANet consisted of four computers located at UCLA, SRI (Palo Alto), UCSB, and Utah. While extremely modest in terms of today's Internet, this was enough to get a small group of pioneers to start thinking about governance and a user's group. So, in 1969 Jon Postel started a list of ARPANet users, which eventually become the telephone directory of the Internet—the *DNS* (Domain Name System).[3] If you wanted to use the ARPANet, you had to ask Postel for a name and address in cyberspace. Once your name and address was entered into the DNS, you became "known" to everyone on the network. Postel's handwritten DNS was soon automated and is now the heart of the Internet.

6.1.1 The Original Sin

The basic DNA of the Internet is a protocol called TCP/IP. Every device connected to the Internet is assigned a number called its IP address. For example, 123.45.67.890 is an IP address of some device somewhere in the world. But humans are not very good with numbers, so the universal resource locator (URL) is a more meaningful name such as Amazon.com or Name@mycompany.us. Before a URL can be encoded within a message it must be converted into a number. This is the job of the DNS—a hierarchical system of servers that map URLs to IP addresses. Every message contains the numerical IP address of source or destination device. TCP is a set of rules called a protocol for accessing machines and transmitting messages from device to device. TCP is a protocol, while IP is a numerical address.

TCP/IP was never designed to be secure; hence the first vulnerability of significance is in TCP/IP itself. TCP/IP is a simple protocol, influenced by the telegraph and adapted to packet switching. Information is packaged in chunks called packets with a source and destination header. By default, the source and destination addresses—such as www.DHS.gov or www.Amazon.com—are in the clear, meaning they can be hacked (IPv4). For example, a malicious user of the Internet can alter the contents of a packet to make it look like email was sent from www.Whitehouse.gov instead of www.TedsSteakHouse.com. This is called *spoofing* and may lead to online fraud.

The TCP/IP protocol uses a very simple handshake to establish a communication link between two computers. When computer A wants to communicate with computer B, it must send a request. When computer B receives the request, it puts it in a list and returns a reply that says, "OK, send your data to me, now," and then waits for the data to arrive. Meanwhile, computer A is supposed to start sending data to B. But what happens if the message never arrives? Furthermore, what if 10 million computers do the same thing to computer B? Computer B is required to save each of the 10 million requests in its memory and wait for messages that never arrive. Ultimately, this causes computer B to run out of memory and shut down. The *SYN flood exploit* is one of the oldest denial-of-service (DoS) attacks known in computing.

The Internet is fragile at the very lowest level of its infrastructure. The code for simple spoofing and DoS attacks can be downloaded from the Internet itself and used to damage it. These weaknesses are in the design of TCP/IP itself. It is as if the human DNA was wired to accept cancer without an immune system to block the malicious cells.

Higher-order structures like email and the WWW are layered on top of TCP/IP. Each of these layers has its own weaknesses, which add to the list of Internet vulnerabilities. At the highest layer, the Internet is composed of major components called *AS* or Tier-1 ISPs that carry most of the traffic and hence pose a high-level network-wide vulnerability.[4] Analysis of major AS shows the Internet's extreme self-organized criticality because of *competitive exclusion* and *preferential attachment*. Recall that competitive exclusion is the force that leads to structures containing a monopoly or oligopoly. There is only one Amazon.com, one Facebook.com, and one Netflix.com of any significance. Preferential attachment is the result of percolation that favors the most popular or efficient node in the network.

As a result, the Internet is a *scale-free network* with very large hubs and critical links that accelerate the spread

[3]Paul Mockapetris[2] of USC/ISI invents DNS with the help of Jon Postel and Craig Partridge in 1983.

[4]An autonomous system (AS) is a collection of Internet routers, switches, and servers under a single administrative control, such as an ISP.

of malicious software such as viruses and worms. As shown in the next section, spectral radius—a measure of self-organization—increases as we incorporate more AS-level nodes in the global Internet. And high spectral radius means high-speed spreading of malware.

The combination of an underlying monoculture (TCP/IP and other standards) that is driving convergence—the movement toward this single monoculture for all forms of communication—and the highly percolated AS network is making the Internet more fragile as its self-organized criticality rises exponentially. We know from biological systems that monocultures, self-organized organisms, and highly percolated complex systems are headed toward major collapse. Thus, the Internet Age is also an age of black swans with chaotic aftershocks. We see this almost daily in terms of massive breaches, rapidly spreading misinformation, and Internet outages.

6.1.2 How TCP/IP Works

The heart of the Internet is a collection of hierarchically organized servers called the DNS servers (see Fig. 6.1). These computers are the Internet's global telephone book. They convert a URL such as Name@Company.com into an IP number such as 123.45.67.890. Every access via TCP/IP must be translated from the URL form to the numerical IP form before a message or access is performed. Thus, the DNS is a hierarchical network of servers as shown in Figure 6.1.

Intermediate-level computers called switches and routers manage the flow of TCP/IP packets through the Internet network. Each switch reads the destination IP address of packets to determine the best route going forward such that the packet is closer to its destination. The router may attempt to find the shortest path that is not too busy, or it may try to find the least expensive path forward. OSPF is the default protocol for forwarding packets from one switch or router to another. Every switch and router contains a list of other routers that it can connect to and forward packets. (This turns out to be a major vulnerability, too.)

TCP is embedded in a bulk message protocol called BGP. BGP collects messages in bulk and transmits the bulk long distances just like the Interstate Highway System handles long-haul freight trucks. BGP is the protocol for the AS-level Internet.

TCP/IP version 4 (IPv4) was officially released in 1976, and its vulnerabilities ultimately led to its replacement, IPv6, first proposed in 1998, but not officially adopted until 2017! By 2020 Ipv4 was still being used by the majority of connected devices. Unfortunately, IPv4 is extremely vulnerable due to its open architecture. Ipv6 is more secure because it provides for encrypted packets and concealment of the path taken by packets as they hop from one server to another on their way to their destination.

UDP, invented by David Reed in 1980, is an alternative to TCP/IP that drops error detection and correction in favor of speed. It is used for streaming audio and video, where a dropped packed is tolerated, but sluggish performance is not.

To show how TCP/IP works, suppose we follow an email message with an attachment, as it goes from a sender in the upper left-hand corner to a receiver in the lower right-hand corner of Figure 6.2. An ATM switch is a very large and powerful router that handles billions of packets per second along the backbone BGP network. The message is chopped into packets and formatted according to a number of protocols and routed through a variety of switches as each packet finds its way through the network. The packets may arrive in any order at the receiver, where they are put back into order and presented to the recipient:

> Step 1: Encoding the message: The email attachment is "wrapped" inside of a MIME (Multi-Purpose Internet Mail Extension) formatted message so that it is recognized as an attachment. The SMTP (Simple Mail Transport Protocol) rule defines how the email and its attachments are handled as they travel through the network. It is important to tag both message and attachment, because they may contain text,

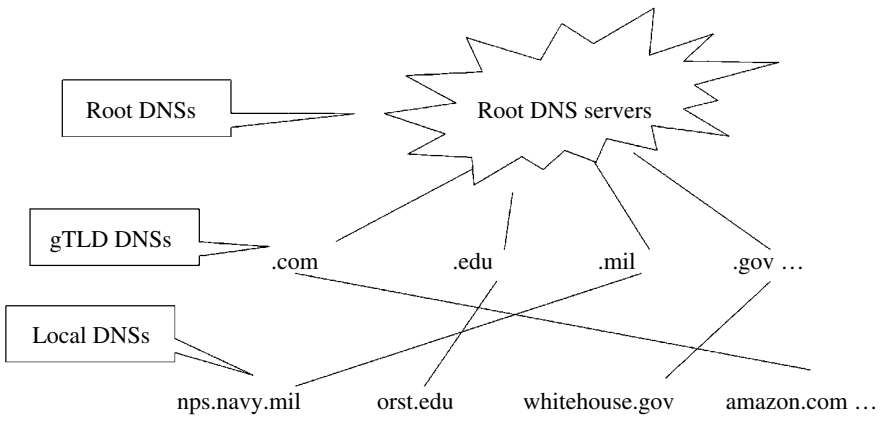


FIGURE 6.1 The DNS is a tree-shaped network of Internet usernames and numbers.

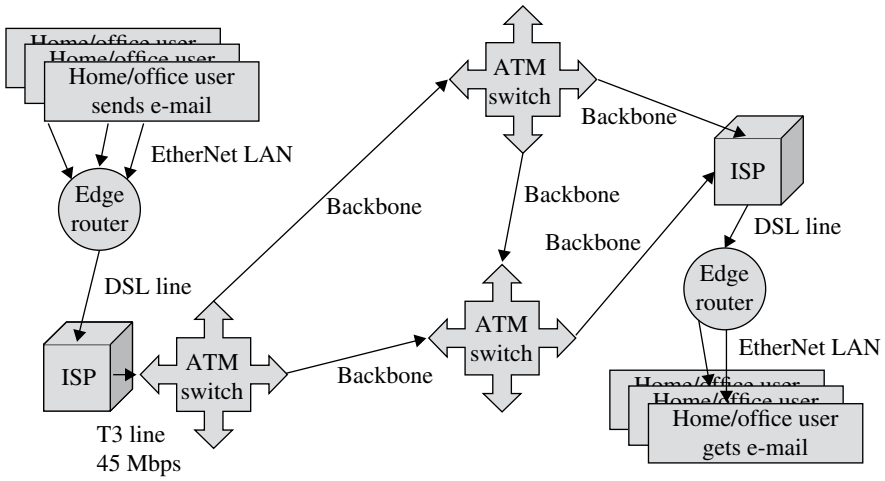


FIGURE 6.2 Example of an email message as it travels through the Internet.

pictures, audio and video information. Each of these types must be handled differently on the receiving end.

Step 2: The sender's machine is connected to the Internet through an *edge router* and local area network (LAN) that communicates in TCP/IP—or UDP if the data is streaming audio or video. So the email and its attachment must be further broken down into packets—TCP/IP or UDP packets. These packets encapsulate the data as well as the source and destination addresses of sender and receiver. But these addresses are in the form of Myname@mymachine.mydomain, instead of an IP address such as 131.200.13.2. So the nearest DNS "telephone book" is consulted to translate the symbolic address from Myname@mymachine.mydomain into 131.200.13.2. Now each packet can be given an address and sent out through the edge router to an ISP (Internet Service Provider) or another switch or router.

Step 3: The ISP provides an onramp to the faster backbone of the Internet. High-speed lines (45 Mbps) and fast switches like ATMs (Asynchronous Transfer Modes) route the email as individual packets along the backbone of the Internet. They use the OSPF (Open Shortest Path First) and BGP (Border Gateway Protocol) protocols to select which routes and physical lines to use. The ISP is an autonomous system (AS) that manages a group of switches and routers. The email packets must cross borders—from the AS managed by the sender's ISP to the AS managed by the receiver's ISP. The BGP rules govern this "border crossing" and routing of the email. In addition, OSPF does exactly what its name implies: using tables stored in the Internet's ATM switches, it selects the shortest available path first. If the shortest route changes in midstream, one packet may take a different path than another. In fact, the packets from the sender's email message may arrive in different order because they took different paths through the network. If TCP was used, then TCP puts the out-of-order packets together again at the receiving end. If UDP was used, the out-of-order packets are discarded. In addition, if a packet is lost, TCP demands that it be sent again, which delays the message, but assures that the entire email message arrives intact.

Step 4: The packets travel across fast lines and ATM switches as they work their way across the network toward the recipient. Switches and transmission lines need to be maintained just like any other physical equipment. But the switches and routers are manufactured by different companies and may work in different ways. SNMP (System Network Management Protocol) is an agreement among all manufacturers on how their devices will be managed. SNMP uses UDP to query and modify the behavior of every device in the Internet. SNMP is the Internet's "in-band" SCADA network.[5] If something goes wrong, an SNMP agent signals this error condition so that the network operation center can take corrective action. Without SNMP, various devices from miscellaneous vendors would not work together, leading to interoperability chaos.

Step 5: The packets arrive at the recipient's desktop and are assembled into proper order according to the rules of TCP. Then the assembly process works its way up the ISO/OSI "stack" (see Fig. 6.5). The SMTP and MIME protocols are worked in reverse order. The TCP packets are grouped into strings of HTML, XML, or pure text. Images and sound are tagged so that an application can recognize them as such. As the email is reconstructed and tagged, it is stored on the recipient's disk drive as a formatted file. Clicking on it causes the appropriate application to open and read the message in the correct format. Note that the email and its attachment can be anything—data, programs, attachments containing audio, video, and pictures. In fact, the attachment can be a malicious program designed to exploit your open system.

This example illustrates the use of routers and switches. Generally, switches (Layer 4: TCP) move packets between routers and switches—typically backbone networks. Routers (Layer 3: IP) move packets between local area routers and switches—typically within a LAN (see Fig. 6.5). Switches are faster and more expensive and so they are used more in

[5]Supervisory Control and Data Acquisition.

backbones. Routers are slower but cheaper, so they are used more in LANs. Routers and switches are managed from a distance using the SNMP and network operation center software.

6.1.3 More Original Sin

TCP/IP and UDP are routed through a subnetwork of switches and routers that users never see, but system administrators are keenly aware of them, because they are the wiring that holds the Internet together. Unfortunately, routers and switches open the door to another original sin of the Internet—tampering with the forwarding IP address tables stored in every router and switch.

Router tables are established by system administrators and network control operators. They are protected by passwords, which are accessed via the SNMP (Simple Network Management Protocol), also accessible to everyone on the Internet. An unspoken honor system among administrators is all that separates secure from non-secure routing. Unfortunately, this system has failed many times. Entire movies have been stolen by re-routing, and re-routing is a known favorite technique of authoritarian regimes.

A North Korean hacker named Park Jin Hyok used re-routing and other techniques to unleash the *WannaCry ransomware* into the Internet in 2017. He was responsible for the Bangladesh Central Bank cyber heist of 2016, Sony Pictures breach of 2014, and other bank and media exploits between 2015 and 2018.

Hyok exploited the fundamental structure of the Internet through fundamental flaws in the design of the TCP/IP and other protocols. IPv4 source and destination addresses are in the clear and can be altered. Routing tables can be changed to force traffic in the AS-level BGP backbone to run through China or North Korea. DoS exploits can bring down DNS servers and fraudulent users can register and join social networks.

The two major vulnerabilities of the Internet are its DNS servers and routers containing routing tables. Access to the DNS servers is highly guarded, but as revealed in Chapter 7, the DNS has been successfully breached as well as the routing tables of switches and routers. Control of the DNS servers and routers essentially gives up control of the Internet.

At a fundamental level, the DNS and routing tables are the Achilles heel of the Internet.

6.2 ANALYZING THE AUTONOMOUS SYSTEM NETWORK

The highest level of the Internet is organized into owner/operator service providers called *AS*. An AS is roughly defined as a collection of Internet routers, switches, and servers under a single administrative control. A single AS may contain a single server or thousands of servers. It may connect to the Internet through one or thousands of connections. AS are numbered from one to over 4 million. To locate an AS, simply search the Internet using its number as a keyword, for example, "AS1 or AS174."

AS are major hubs in a hub-and-spoke architecture that forms a dense network. The nodes of this network are the AS service providers and the links are the *peering* routes connecting to other AS providers. These links may be uploads or downloads, or both. We will not be concerned with the direction of the links, however. Generally, all regions of the Internet are highly *percolated*, which suggests that the Internet is highly self-organized. This is illustrated by the following study of the top AS nodes in the global Internet.

6.2.1 The AS500 Network

Figure 6.3 shows the network formed by linking the 500 largest AS in the Internet with their peers through 4564 routes. Recall that the 1996 Telecommunications Act required ISPs to peer—meaning they must interoperate and allow one another to rent each other's networks. Peering leads to globalization of the Internet by linking together independently owned and operated TCP/IP networks.

Figure 6.3 shows the so-called route list available through the Cooperative Association for Internet Data Analysis (CAIDA) circa 2004. Overall, the CAIDA route list contained 42,000 AS and over 121,000 routes in 2004.[6] An analysis of all 42,000 nodes is beyond the scope of this book. However, the fragility of the Internet relative to the spread of malicious software is easily revealed by a study of the 500, 1000, and 2000 most connected AS nodes. (The 2020 route list is even much larger.)

The most connected nodes of the AS-level network shown in Figure 6.4 was created by searching all 42,000 known AS and deleting the least connected nodes. In this way, the remaining 500 nodes and 4564 links form an Internet backbone, sufficient to study its resilience. The mean connectivity of the 500-node AS network is 3.42 connections and its spectral radius is 14.9—over four times larger. This says the AS network is modestly dense and highly structured. Accordingly, it is perched on the edge of self-organized criticality. Self-organized criticality means malware travels fast and far, and resilience is low.

The top 10 AS nodes in Figure 6.3, and their degree of connectivity, are:

AS701: Verizon—91
AS721: DOD—79
AS1239: Sprint—55

[6]http://www.caida.org

AS1: Level3—50
AS209: Qwest—48
AS286: KPN—40 (Trans-Atlantic)
AS293: ESNet—40 (Energy Sciences Network—US Nat'l Labs)
AS702: Verizon—33
AS2516: KDDI—29 (Japan)
AS1913: DOD—28

Similarly, the largest betweener routes are:

AS701 → AS715
AS701 → AS721
AS701 → AS702
AS721 → AS765
AS701 → AS714
AS721 → AS745

The largest betweener—AS701—has 91,552 paths running through it. Clearly, the competitive exclusion principle is at work in this sector, creating a handful of super-connected hubs. The largest AS operators, known as Tier-1 ISPs, dominate the Internet. In 2013 there were only six Tier-1 ISPs: Level 3 Communications, CenturyLink, Cable & Wireless Worldwide, UUNet, Sprint, AT&T Corporation, and Genuity.[7] A decade earlier, there were more than 100. Gause's law ran rampant on the Internet when it was subject to very little regulation and economics dictated centralization.The AS-level network is highly percolated and structured, as indicated by its spectral radius, which is many times larger than the mean connectivity in AS500, as well as larger versions of the AS-level network. For example, the spectral radius for 500-, 1000-, and 2000-node AS-level networks steadily increases as the size of the core increases:

500 nodes: 14.9
1000 nodes: 29.7
2000 nodes: 63.9

The Internet is vulnerable to malicious software exploits such as viruses or worms even with very low levels of infectiousness. Malware that jumps from one computer to another with probability less than 1% is highly likely to spread to most of the global Internet without dying out. This is why malware has been highly successful as a hazard of the open Internet.

The reason for such high risk of infection in the global Internet is shown graphically in Figure 6.4. This core network was reproduced from the largest hubs and betweeners extracted from Figure 6.3. Local network neighborhoods are

[7]https://en.wikipedia.org/wiki/Internet_backbone

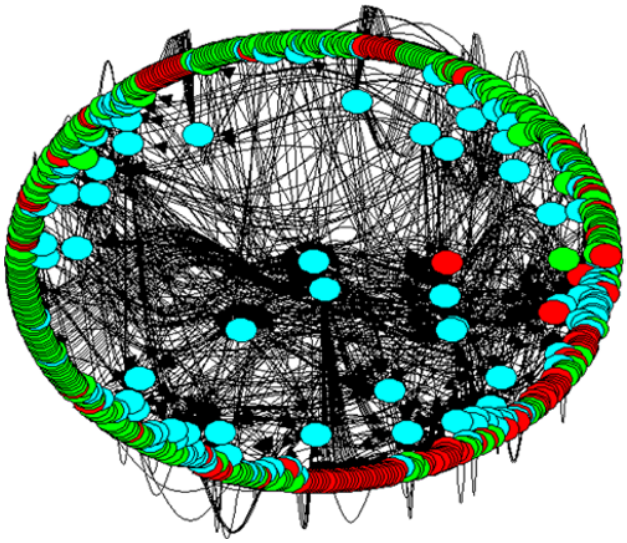


FIGURE 6.3 The top 500 autonomous system servers in the Internet, circa 2004, arranged here according to their connectivity. Nodes with degree of connectivity are placed in the center, and low-connectivity nodes are placed around the circumference.

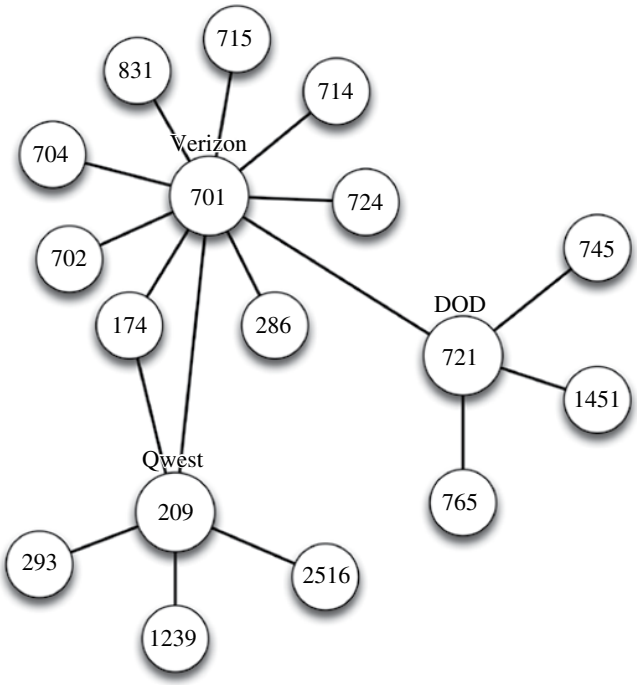


FIGURE 6.4 Core of the AS500 Internet circa 2004 contained the most connected nodes and the highest betweenness links.

characterized by hubs with relatively high connectivity through high-betweener-valued links. It is easy to infect the Internet because hubs are super-spreaders of viruses and worms and high-betweenness pathways are easily traversed.

Resilience against the spread of viruses can be achieved by reducing the size of these hubs or by hardening them

against malicious software attacks. The likelihood of a malicious exploit passing through a hub or betweener link is much higher than through other nodes and links. Therefore, blocking the spread of such exploits at the hubs and betweeners is more effective than blocking at the desktop or handset level. Protection of individual PC, cell phones, and tablets is far less effective than protection at the large AS-level hubs and links. The best place to protect users is to protect the AS-level super-spreaders.

Similar conclusions can be drawn from an analysis of robustness. The Internet is extremely robust against link and node de-percolation. For the AS500 network with mean connectivity of 3.42 and spectral radius of 14.9, we can expect link and node robustness to be relatively high:

$$\kappa_L = 1 - 2/\lambda = 42\%$$
$$\kappa_N = 1 - 1/\rho = 93\%$$

An attacker would have to remove 0.42(4564) = 1917 links and select one of the 0.07(500) = 35 critical nodes to separate the AS500 Internet into islands. Which links are critical? The only way to determine this is by trial and error. On the other hand, the most connected nodes are the most likely to separate the Internet into disjoint components. Therefore, major hubs are the most critical nodes and the most easily identified targets.

6.2.2 Countermeasures

Owners and operators are painfully aware of the vulnerability of the Internet and its open protocols to malware. They have a wide array of countermeasures to apply to prevent malware from penetrating corporate networks and spreading to other networks. These countermeasures are described in greater detail in the next two chapters. This section describes Internet-wide countermeasures and safe practices at a high level.

Malware travels through the Internet via ports. Every computer and cell phone has one or more ports for input and output. Ports are numbed and standardized so that anyone can use them. For example, port 80 is the unencrypted port for browsing. Port 443 is a Web browser's secure port. There are potentially thousands of ports for exchanging data between pairs of devices.

Ports are also doorways for malware. In fact, ports are the main way malware spreads from computer to computer and IoT device to IoT device. Protecting ports is the major responsibility of owners and operators of handheld, desktop, enterprise server, and cloud computing systems.

Ports are managed by physical and software-defined firewalls. Every computer has a firewall. Firewalls contain white lists (lists of open ports) and black lists (lists of blocked or closed ports). Most ports should be closed on most computers.

Intrusion detection systems prove additional protection by analyzing packets as they enter a computer system. A good intrusion detection system can identify the fingerprints of known malware and divert DoS attacks. They are limited, however, unless backed up by sophisticated software for identifying malware and adapting to constantly changing malware.

The DNS is the heart of the Internet. Physical access to the 13 root DNS servers is protected and cyber access is constantly monitored. Access is routinely allowed so that URLs can be quickly translated into IP addresses. But this process is done under tight control. In particular a cryptographic handshake is required to complete a translation. A request must be cryptographically signed to authenticate the requester, and the DNS must handle cryptographic certificates to consummate the transaction. Public key infrastructure (PKI) is described in detail in Chapter 8.

A number of "dark web" enhancements have been proposed and implemented to conceal the identity of users while allowing free access. The Onion Router (TOR) is one such enhancement. "Tor is based on the principle of 'onion routing', which was developed by Paul Syverson, Michael G. Reed and David Goldschlag at the United States Naval Research Laboratory in the 1990's. The alpha version of Tor, named 'The Onion Routing Project' or simply TOR Project, was developed by Roger Dingledine and Nick Mathewson. It was launched on September 20, 2002. Further development was carried under the financial roof of the Electronic Frontier Foundation (EFF)."[8] The TOR browser is a Web browser that encrypts IP addresses used to browse the Web.

TOR works as follows: Transactions between a user and a Web site bounce from random site to randomly selected site until reaching a final destination site. At each waypoint, the next waypoint is selected at random until the final destination is selected. Additionally, the IP addresses of waypoints are encrypted and nested much like the layers of an onion. Each waypoint unwraps the encrypted next IP address, decodes it, and forwards the message to the next waypoint, where it is unwrapped again, until reaching the end of a chain of waypoints.

IPv6 provides support for VPN technology that works similar to TOR, but without randomization. The source and destination of IPv6 packets is encrypted to prevent man-in-the-middle exploits. The VPN consists of VPN-enabled servers embedded in the public Internet. When a VPN packet arrives at a VPM-enabled server, its source and destination addresses are decoded, and the next destination address encrypted and substituted into the packet so it can be forwarded to another VPN-enabled server. VPNs rely on dedicated VPN-enabled servers that overlay on top of the open Internet.

[8]https://fossbytes.com/everything-tor-tor-tor-works/

Blockchain is another proposed enhancement to secure DNS translations. A blockchain is a ratcheted chain of encrypted links. Each block in the chain contains an encrypted pointer to the next block. Much like the TOR router, the chain of blocks is encrypted and must be decoded in order, from beginning to end. A blockchain is a *ratchet* that works only one way—from beginning to end. Unlike TOR, blockchains are distributed over a peer-to-peer network of nodes, each node containing a copy of the chain. The distributed blockchain adds redundancy and consensus to authentication.

One proposed application of blockchain to the operation of the DNS is to treat the DNS as a distributed blockchain. This increases resilience under stress or loss of one or more nodes. It also provides a form of security because a majority of nodes must agree on the authenticity of requests and accesses before they are allowed. Blockchains and onion routers both use public and private key encryption to secure transactions. Public and private key infrastructure is described in detail in Chapter 8.

6.3 THE RFC PROCESS

Internet standardization started appearing very early in the history of the Internet. Steve Crocker of UCLA created a public process called RFC, which became the major tool of Internet self-organization and decision-making. RFC 1 was issued by Steve Crocker on April 7, 1969, and describes the first Internet switch—called *IMP* (Interface Message Processor). Back then modifications to the Internet were all vetted through an RFC. For example, RFC 688 documented a new standard for email in 1975. By 2004, there were over 3700 RFCs on record.[9]

The RFC process is an example of an emergent process that starts at the bottom and works up to the top. This is in sharp contrast to the way most governing bodies work. Generally a governing body dictates from the top down. Two key technologies demonstrate the emergent RFC process: email and TCP/IP.

6.3.1 Emergence of Email

In 1971–1972, Ray Tomlinson invented what we now know of as *email*, and Larry Roberts quickly improved on it. Tomlinson started using the "@" character to separate username from computer and domain name in his email headers, for example, name@machine.com. This convention soon became the standard method of addressing email:

> The first message was sent between two machines that were literally side by side. The only physical connection they had (aside from the floor they sat on) was through the ARPANET. I sent a number of test messages to myself from one machine to the other. The test messages were entirely forgettable and I have, therefore, forgotten them. Most likely the first message was QWERTYUIOP or something similar.[10]

Today's email system is more complicated because it includes the ability to embed pictures and sound, and it handles attachments, which increase the vulnerability of the Internet to hacking. The original email standard based on RFC 822 (1982) was replaced by RFC 2822 in 2001. Email evolved through the RFC process into what we know as email today.

6.3.2 Emergence of TCP/IP

A seminal event took place in 1973 that marks the technical beginning of the modern Internet. Vinton Cerf of Stanford and Robert Kahn of DARPA invented TCP (Transmission Control Program)—to put packet-based communications on a solid and reliable footing. The term "Internet" and the TCP/IP protocol that emerged circa 1973–1976 were created spontaneously and simultaneously. By 1976, DARPA required the use of TCP in ARPANet, and the Internet was officially operational. This evolution is documented in a series of RFCs. Today's TCP/IP protocol is defined by RFC791 (1981) and explained in a tutorial given by RFC1180.

TCP is the sequencing half of the TCP/IP protocol. Its job is to reassemble packets after they arrive at their destination. An email message, for example, might consist of thousands of packets. Each packet may take a different route from source to destination and therefore arrive out of order. TCP puts them back in order and checks to make sure there are no errors. (If an error is found, TCP insists on a retransmission.)

Also during this period, Robert Metcalfe at Xerox PARC was working on a protocol for connecting computers together over a local area network (LAN). His solution led to the invention of *Ethernet* for LANs. This "local area network protocol" became the dominant LAN protocol and IEEE 802.3 standard. Eventually, the novel ideas in Ethernet became the basis for IP—the other half of TCP/IP.

Ethernet was a significant advance because it overcame a limitation of TCP. When two or more computers attempt to send a message at the same time, an electronic *collision* occurs. Both messages are garbled because only one message can be transmitted at a time. Metcalfe proposed an elegant solution. Whenever a collision occurs, both computers try transmitting again after waiting a random length of time. Because the probability of the two computers waiting the same random length of time approaches zero as the number

[9]http://www.rfc-archive.org/

[10]Tomlinson, Ray, *The First Network Email*, http://openmap.bbn.com/~tomlinso/ray/firstemailframe.html

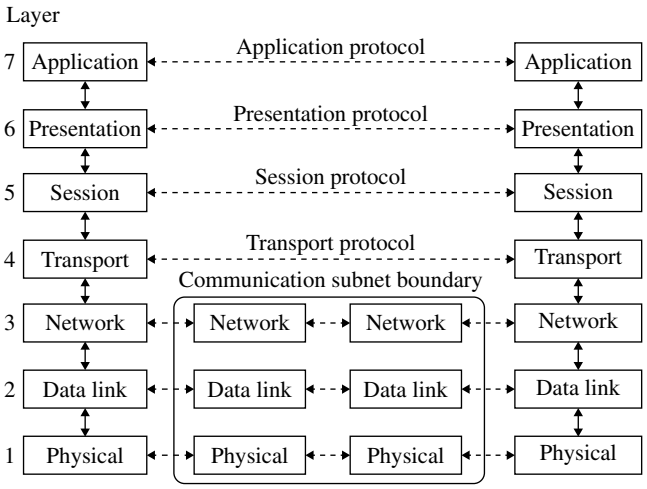


FIGURE 6.5 The ISO/OSI standard protocol stack for the Internet consists of eight layers.

of re-try's increases, both computers eventually succeed in sending their messages.[11]

TCP was not perfect—and still is not. But even as early as 1978 Vinton Cerf, Jon Postel, and Danny Cohen realized that TCP was trying to do too much. It was too big. So they decided to divide TCP into two parts: TCP and IP. Thus TCP/IP was born. As the protocol took on more functionality, it became necessary to further divide it and put different functions into different layers. TCP and IP are two of the seven layers that define the IP today (see Fig. 6.5).

The first layer of the IP (Layer 1: Physical) consists of a wire, optical cable, or some other physical device. The second layer (Layer 2: Data Link) defines the packet format and how to deal with collisions. Layer 2 is essentially Ethernet. The third layer (Layer 3: Internet Protocol) is the "IP" part of TCP/IP and defines how packets are routed from sender to receiver. Inside of every Internet switch is a routing table that tells each IP packet where to go next. The next layer controls how IP packets are recovered at the other end. Layer 4: Transport can be implemented as one of two protocols: TCP or UDP. TCP guarantees delivery of all packets and reorders any packet that arrives out of order. It keeps track of the packet delivery order and the packets that must be resent. UDP is faster, but less reliable as it does not guarantee delivery and does not bother to reorder packets. UDP can lose packets.

TCP is used for email and most Internet transmissions, and UDP is used for streaming media such as video and audio, where a missing packet or two will not be noticed. UDP is fast because it does not have to reorder or retransmit packets. But UDP is less reliable.

The four-layer TCP/IP protocol described above defines the modern Internet. In 1988 the International Standards Organization released the Open Systems Interconnection (OSI) standard—a competitor to TCP/IP. OSI defines three more layers: Layer 5: Session, Layer 6: Presentation, and Layer 7: Application. Figure 6.5 shows all layers of the ISO/OSI standard. While ISO/OSI is the international standard, the popularity of PCs running TCP/IP, and the fact that most of the servers on the Internet run TCP/IP, means that the Internet is for all practical purposes, identical to the four-layer TCP/IP protocol.

TCP/IP emerged. It was not invented in one big bang. Rather, it was the by-product of the RFC process.

TCP/IP is the basic infrastructure upon which the entire Internet and WWW depends on. It is at the heart of the Internet's DNA. Unfortunately, it is extremely vulnerable to many exploits ranging from simple DoS to malicious code disguised as an email attachment. TCP/IP was never designed to be secure, and in fact the opposite is true—it was designed to be simple and open. The lingua franca of the Internet is the first weak link in the global communications network we call the Internet.

6.4 THE INTERNET OF THINGS (IOT)

The IoT encompasses what Mark Weiser (1952–1999) called *ubiquitous computing* in 1988. Weiser noted that almost anything—clothing, tools, appliances, cars, and coffee mugs—could be connected to the global Internet. The number of connected things rises exponentially due to the exponentially rising Moore's law that was in full swing from 1965 to 2015. As the Internet expanded at an exponential rate following its commercialization in 1998, computing became pervasive, according to Uwe Hansmann of IBM. He noted that technology was moving beyond the PC to everyday devices. Hansmann rebranded ubiquitous computing as *pervasive computing* and characterized it as "computing devices becoming progressively smaller and more powerful."

Kevin Ashton coined the phrase *the Internet of Things* to describe the network connecting objects in the physical world to the Internet.[12] His definition of ubiquitous computing incorporated sensors as well as progressively small and more powerful processors. Sensors collect and exchange data obtained from their environment. Subsequently, IoT has become a major factor in computer security because of the threat of collection and processing of unauthorized data:

$$IoT = \text{ubiquity} + \text{processing power} + \text{sensors} + \text{connectivity}$$

Billions of computers embedded in billions of everyday devices are collecting and aggregating all kinds of information

[11]This is called Carrier Sense Multiple Access/Collision Detection (CSMA/CD).

[12]http://www.smithsonianmag.com/innovation/kevin-ashton-describes-the-internet-of-things-180953749/#S6Smb6YwsPhm17uW.99

on the location, behavior, activity, buying habits, and social connections of billions of people across the globe. IoT collects and aggregates where individuals are (home, shopping mall, school, work), what people are doing (running, talking, driving, sitting, eating, shopping), and personal behaviors (biorhythms, habits, heartbeats, and automobile driving behaviors). Devices such as the Amazon.com Echo listen to everything you say to aggregate what it hears and target consumers with advertisements that anticipate future wants and desires.

6.4.1 Data Scraping

Individual IoT devices may collect relatively narrow personal information such as a person's heartbeat and location, but when combined with other sensor-derived data and interpreted by automated processes such as machine learning algorithms, an Internet organization can build a detailed profile of every person on the planet. The Internet organization may be friendly or not, depending on its purpose. A friendly organization might build and use consumer profiles to anticipate future purchases, such as what online movies the consumer might want to watch next. A malicious organization might use a consumer's profile to derive passwords or empty bank accounts.

Internet companies like Google.com, Facebook.com, and Netflix.com collect data on their users to tailor searches, target advertisements, sell consumer profiles to other companies, and recommend future purchases. For example, Netflix.com uses profiles to push movies it knows targeted users will like. It also aggregates profiles so it can manufacture new movies with known audience demographics. The company knows that images with fewer than three people in them are better liked than images with more than three people; villains are preferred over heroes; complicated expressions on people's faces are preferred over stoic or benign expressions. The movie *Orange Is the New Black* was manufactured from aggregated data obtained from Netflix.com consumers.

When combined with machine learning algorithms, data scraping from IoT devices opens the door to malevolent organizations and machines to "take control." It enables autonomous control that leaves humans out of the loop, such as found in self-driving cars, or automated loan approvals based on how a person seeking a loan uses his or her smartphone. It allows for big data analytics to bypass privacy and security controls. For example, Facebook.com user passwords can be mined with 90% success rate by merely analyzing a user's Facebook.com page and postings. When users are not informed of what data is collected and how it is used, control passes from consumer to organization. When the organization makes the decision based on IoT data, the human consumer is left in the dark.

6.4.2 IoT Devices

Many IoT devices prone to exploitation by malicious organizations are so commonplace their misuse may be overlooked by consumers. The health information collected by a digital watch or video collected by a security camera at home may not seem like a threat to privacy, but when combined with other personal information, privacy is no longer assured. IoT devices have become so commonplace that we often forget their potential for abuse. Many IoT devices hide in the background and consumers are not aware of them.

The following list is a subset of the devices known to be compromised by malicious actors. They are both obvious and not-so-obvious collectors of personal and private data. Refrigerators, smart TVs, digital watches, webcams on home security systems and desktop computers, smartphones with various health sensors and GPS, printers, ATM bank machines, gasoline pumps, industrial controls, automobiles, traffic lights, and health monitoring devices are common examples. This list continues to get longer as pervasive computing gets more pervasive.

A printer at the University of Michigan was infected by malware that began generating DoS attacks against other institutions. The source of the DoS was difficult to find because printers are not known to attack people. Similarly, a water treatment facility in Michigan was hacked when its simple password was cracked by malware. Most SCADA control systems are delivered with simple passwords, such as "password" or "123456," and never changed. Table 6.1 lists

TABLE 6.1 A sampling of the most common usernames and passwords that are rarely changed, but frequently get hacked by malicious organizations

User name	Password
(none)	(none)
adm	(none)
admin	password
Cisco	Cisco
customer	(none)
debug	synnet
device	device
guest	(none)
HELLO	FIELD.SUPPORT
manager	manager
manager	friend
monitor	monitor
OPERATOR	SYS
root	(none)
security	security
setup	setup
tech	tech
User	(none)

Source: http://www.phenoelit.org/dpl/dpl.html.

the top username and passwords used by scanning Web site shodan.io, an IoT search engine released by Internet cartographer John Matherly in 2009. It shows how some owners and operators of highly critical infrastructure systems are easily hacked due to carelessness and incompetence.

BASHLITE is an IoT DDoS malware that carries with it a built-in list of common usernames and passwords. In many cases, IoT devices are shipped from the manufacturer without a meaningful username or password. Even if the password is randomized, the same password may be used in all devices shipped over a long period of time. Discovery of the default password puts all devices at risk. In 2019, the State of California passed a law requiring every IoT device be delivered with a unique password.

Gasoline pumps, airline entertainment systems, boarding pass machines, pharmacy laboratory equipment, taxi cab kiosks, and X-ray tomography machines have been infected by IoT malware. The Mirai virus used a list of 60 common default usernames and passwords to install a DDoS botnet across Linux-based IoT devices in 2016. Derivatives based on Mirai persist to this day, as hackers modify and improve its virality.

Heart pacemakers, insulin pumps, and defibrillators have been shown to be susceptible to exploitation by malware, although attacks on these devices have not been widespread. Theoretically, malware can stop pacemakers, overdose patients with insulin, and shock defibrillator patients to death. For some products, the only thing malware needs to know is the six-digit serial number on the device. Hacking an insulin pump via wireless connection from within 50 ft of a patient was demonstrated in 2011.

Jackpotting is the process of forcing automated teller machines (ATMs) to dispense cash without withdrawal from a bank account. Jack Barnaby demonstrated Jackpotting at a conference in 2010. Carjacking is the process of hacking the electronics of an automobile. Every modern automobile contains a local network and hundreds of computers that control entertainment systems, dashboard functions, steering, braking, and transmission gears. Carjacking is performed by gaining access to the car's *telematics* system based on the ODB (onboard diagnostic) network. In particular, the federally mandated OBD-II (onboard diagnostics) port under the dash is the most common gateway into the car area network (CAN). Once compromised, malware can take control of the car through OBD-II using SSH (Secure Shell) commands.

Driver fingerprinting is an example of invading the privacy of drivers using machine learning techniques on aggregated data. Researchers have shown how fingerprinting is used to identify the driver by simply recording driving behaviors. In a test of 15 drivers performing maneuvers in a parking lot and along a 50-mile loop through Seattle, Washington, a machine learning algorithm identified drivers with 100% accuracy using only 8 min of collected data.

6.4.3 More IoT Exploits

As the number of devices connected to the Internet increases, so does the number and variety of exploits. The following list of known exploits is likely surpassed by more recent exploits since the 2020 publication date of this book.

The Mirai DDoS attack of 2016 infected large portions of the global Internet including major e-commerce sites such as Twitter.com, Netflix.com, and Reddit.com. Mirai is Japanese for "future." It principally attacked Linux-based devices such as webcams and home routers using a built-in table of 60 common factory-default usernames and passwords. A number of enhanced derivatives (Satori, Okiru, Masuta, and IoTroop) have appeared since publication of its source code online.

The St. Jude Medical's implantable pacemaker and defibrillator and Owlet Wi-Fi baby heart monitor were shown to be vulnerable to malware in 2016–2017. The St. Jude Medical's device vulnerabilities were patched subsequent to a US CERT advisory. US CERT identifies common vulnerabilities and exposures (CVEs) using a combination of year and exploit number. Specifically, CVE-2017-12712 allows a nearby attacker to issue unauthorized commands to the pacemaker via wireless communication. A second vulnerability, CVE-2017-12714, allows a nearby attacker to drain the device's battery by repeatedly sending commands.

In 2012 the TRENDnet Webcam hack exploited home security cameras due to faulty software that allowed anyone with the webcam's IP address to access its streaming video signal. Unauthorized webcam videos of a Laundromat in Los Angeles, a bar in Virginia, living rooms in private homes, and a man watching a football game and numerous nudity videos were posted on public Web sites without permission. Many of the faulty webcams remained unpatched 5 years after the vulnerability was discovered and a software update made available to consumers.

Researchers performed a Jeep attack in 2015 to illustrate a common vulnerability in automobiles. They hijacked the vehicle over the Sprint cellular network and made it speed up, slow down, and drive off the road. The researchers accessed the OBD-II port using off-the-shelf SSH commands and then issued instructions to the Jeep's onboard computers.

Automobiles are becoming more automated and connected to the global Internet. This connectivity is intended to prevent one automobile from colliding with others and provide autopilot-like control. But when placed in the wrong hands, connectivity opens the door to malicious code that takes complete control of the vehicle.

Additional details on the spread of malware and dangers of botnets and ransomware are presented in subsequent chapters. The foregoing examples illustrate the vulnerability of TCP/IP and other IP that were designed decades ago, before malicious hackers ever dreamed of hacking. They underscore the inherent weakness of the Internet, which can

only be removed by fundamentally different protocols and designs that include security as a requirement.

6.5 COMMERCIALIZATION

In 1981 the NSF established a research network based on TCP/IP called CSNet. It was aimed at serving the non-ARPANet users in the broader academic research community. Business was so good that CSNet was "outsourced" to MCI and renamed NSFNet. Then in 1990, ARPANet merged back with NSFNet! Once again, the Internet was whole—and interoperable with anyone that adhered to the TCP/IP protocol.[13]

Meanwhile, the Internet was increasing in value and popularity. Over one million users paid a subscription fee to login to NSFNet by 1992. A National Research Council report chaired by Leonard Kleinrock suggested that the NSFNet be commercialized (at this time it was still the responsibility of NSF). His report attracted the attention of Vice President Albert Gore and in 1999 the Vice President of the United States claimed parentage of the Internet in an effort to get elected President.[14] In 1992 the US Congress gave NSF permission to commercialize the Internet over a 5-year period of time. This began a 5-year process of transition that ended with the privatization—indeed, the globalization—of the Internet.

A year later (1993), the number of subscribers had doubled to 2 million. The NSF created InterNIC to support the rapidly growing Internet and contracted with AT&T to maintain the DNS structure. In addition, the NSF awarded a 5-year contract to Network Solutions, Inc. to sell domain names for $50/year. During this period, millions of people became subscribers—fueling the *Internet Bubble* that eventually burst in March 2000.[15]

After spending $200 million from 1986 to 1995, NSF outsourced the Internet DNS to four companies and turned the business of doing Internet business over to the US Department of Commerce. In 1997 the Clinton administration directed the Secretary of Commerce to privatize the DNS, "in a manner that increases competition and International participation." True to the Internet culture as Steve Crocker defined it, an RFC-like "white paper" was circulated by the US Department of Commerce. In 1998, the Internet was set free.

[13]Cisco Systems, one of the most successful companies to commercialize TCP/IP equipment, was founded in 1984 by Leonard Bosack and Sandra Lerner, who later sold their interests in the company for $170 million.

[14]During a March 1999 CNN interview, while trying to differentiate himself from rival Bill Bradley, Gore boasted: "During my service in the United States Congress, I took the initiative in creating the Internet."

[15]The Internet Bubble (1995–2000) was a period of economic excess where billions of dollars were invested in "dot-com" startups attempting to commercialize the Internet. A few of these startups survived, for example, Amazon.com, Yahoo.com, and Ebay.com, but most of them went out of business, leaving many stock market speculators stunned.

The rapid growth of the Internet since 1998 has been nothing but phenomenal. *Metcalfe's law* explains the Internet's explosive growth in terms of a *network effect*: the value of a communications network is proportional to the *square* of the number of connected users. Given a network with n nodes, it is possible to connect every node to every other node through $n(n-1)/2$ links. Rounding off, this means a network magnifies connections among users and devices by a factor of n^2. Does Metcalfe's law explain the rapid growth of the world's largest communications system? Not quite.

6.6 THE WORLD WIDE WEB

Progress continued at a rapid rate throughout the 1980s and 1990s as the Internet coevolved with the rise of the low-cost PC. In 1979 there were 100 users of ARPANet. In 1984 the number had grown by a factor of 10–1,000 users, and another factor of 10 brought the total to 100,000 users by 1990! But the number of Internet users would never rival that of radio or TV unless the Internet offered something more than connectivity. What the infant network needed was applications—or better yet, the *killer application*.

In 1982 when Jon Postel established SMTP (RFC821 and now RFC2821) as the standard for doing email, the killer application of the Internet seemed to be email, because most of the data traveling over the Internet were email messages. Even the defense, research, and university communities used the Internet mainly for email. (This was a curious outcome, since the original purpose of the Internet was to share large centrally managed mainframes.)

The killer application for the Internet—the application that would ignite mainstream adoptions of networking—was invented by Tim Berners-Lee while he was working for the world's largest particle physics research laboratory—the Center for European Nuclear Research (CERN). In 1989, Berners-Lee invented the *World Wide Web*—a network of hyperlinked documents accessible via the Internet. Then, he built the first browser and invented HTML (Hypertext Meta-Language) to support the sharing of hyperlinked documents across the Internet. His goal was to simplify the publication of research papers so that any physicist could disseminate his or her research electronically. What if an author could simply imbed a hypertext URL in any text document so that another document could be selected and retrieved merely by clicking on the embedded hyperlink? This would simplify the retrieval of referenced papers, regardless of where they were stored. One document could come from machine A, another document from machine B, and another document from machine C. Regardless of where the document lived, the collection of documents would pop up on the user's screen as if it were part of one large collection.

The hyperlinked document idea was not new. Ted Nelson had proposed hyperlinks a decade earlier. Even so,

Berners-Lee had to overcome the mind-set of the Internet, which was that networking was designed to connect computers to users and users to computers. The bigger the computer, the more users needed the network connection. But Berners-Lee had a better idea. Why not connect users to documents, regardless of where they are? Users wanted information, not connectivity to hardware. This is obvious in hindsight, but at the time, it was a contrarian's view of what the Internet was good for.

Berners-Lee called his software a *browser-editor*, because it combined a text editor with a web of hyperlinked documents. It provided a powerful tool for scientists, but it lacked the ease of use that consumers accustomed to a graphical user interface expected. What the WWW needed was a browser that worked like the graphical user interface on a Macintosh PC. If an ordinary consumer can use a PC with a point-and-click interface, he or she should be able to use the WWW—and this required a simpler interface.

Marc Andreessen and Eric Bina developed the first graphical browser for the WWW while students at the University of Illinois–Urbana. MOSAIC was a better mousetrap because it simplified the user interface. Originally developed on the NeXT workstation, the two students quickly ported it to the Macintosh and Windows PC. The WWW experienced explosive growth when MOSAIC became available for inexpensive and ubiquitous Macintosh and PC computers.

Andreessen's and Bina's invention was more than an easier-to-use browser-editor. It enhanced the hypertext language invented by Berners-Lee in several important ways. According to Andreessen,

> Especially important was the inclusion of the "image" tag which allowed to include images on web pages. Earlier browsers allowed the viewing of pictures, but only as separate files. Mosaic made it possible for images and text to appear on the same page. Mosaic also sported a graphical interface with clickable buttons that let users navigate easily and controls that let users scroll through text with ease. Another innovative feature was the hyper-link. In earlier browsers hypertext links had reference numbers that the user typed in to navigate to the linked document. Hyper-links allowed the user to simply click on a link to retrieve a document.[16]

Andreessen and Bina moved to California, co-founded Netscape Communication Corporation with money from Jim Clark, rewrote MOSAIC, and called it *Netscape Navigator*. The trio built the first Internet Age company on top of the Internet infrastructure—Netscape Communications Company. The highly successful enterprise was sold to AOL in 1999, but for a brief time, it was the fastest-growing company in America.[17] Even more significant, Netscape ignited the commercial Internet. At the time of its public offering in 1995, Netscape claimed 35 million users. Five years later, the Internet had over 250 million users—and 75% of them used Netscape's browser.

A large installed base of PC users, an easy-to-use graphical browser, and a cleverly designed WWW all came together in 1995 to propel the Internet into the mainstream. During the 5-year period from 1995 to 2000, adoption of the Internet far exceeded the 30-year adoption rate of cable TV, 20-year adoption rate of the home computer, and the 15-year adoption rate of the VHS/VCR.[18] The Internet achieved 50% market penetration in 5 years—an adoption rate that has yet to be beat by many other global products.

By the end of the dot-com bubble in 2000, most of the infrastructure we know of as the Internet was in place. Unfortunately, it is based on a TCP/IP monoculture susceptible to cyber exploits and a highly percolated complex system far beyond its self-organized criticality. This highly decentralized, self-organizing system has evolved to a highly fragile ecosystem under highly deregulated and open conditions. And yet, it is no more vulnerable and fragile than the overly regulated and highly fragile energy, power, and transportation systems that evolved under radically different conditions. It seems that self-organization is every complex system's destiny.

6.7 INTERNET GOVERNANCE

A question often asked is, "Who owns the Internet?" Other infrastructure sectors are owned by corporations or jointly by public–private partnerships. Public utilities (water and power) are often pseudo-private, meaning they are either heavily regulated monopolies, or completely owned and operated by a municipality or metropolitan region. The Internet is different, because for one thing, it is a global organization. Its governance resembles the United Nations more than Microsoft Corporation or the federal government. The "UN of cyberspace" is actually a loose collection of societies—mainly run by volunteers. Figure 6.6 lists some of the groups that play a major role in Internet standards, design, and ethics. This is a partial list, but it is the sustaining core that keeps the Internet going and evolving.

6.7.1 IAB and IETF

The Internet is an open society of many volunteer organizations simultaneously contributing new ideas and technical recommendations for its evolution. It is a decentralized, free-wheeling society that evolves standards rather than dictates them. One of the earliest of these voluntary organizations was the Internet Architecture Board (IAB) formed by Barry Leiner. According to RFC1120, the IAB:

[16]http://www.ibiblio.org/pioneers/andreesen.html

[17]In 1999, AOL paid $10 billion in stock for 5-year-old Netscape.

[18]It took 67 years for the public telephone to penetrate 50% of the homes in the United States.

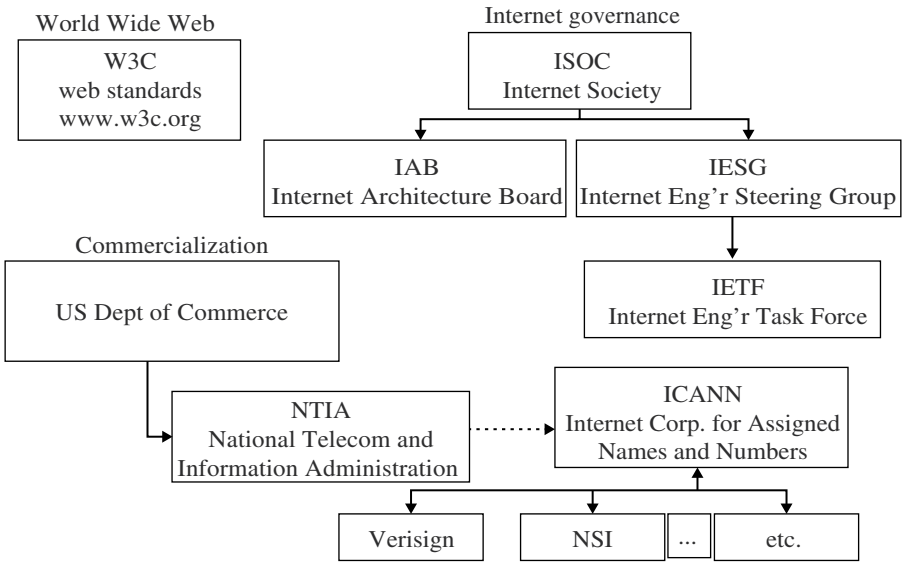


FIGURE 6.6 The core of Internet governance circa 2010 included W3C, ISOC, IETF, ICANN, and related agencies within the United States and partially globally.

1. Sets Internet standards,
2. Manages the RFC publication process,
3. Reviews the operation of the IETF and IRTF,[19]
4. Performs strategic planning for the Internet, identifying long-range problems and opportunities,
5. Acts as a technical policy liaison and representative for the Internet community, and
6. Resolves technical issues which cannot be treated within the IETF or IRTF frameworks.

Perhaps the most significant influence on the Internet has come from the activities of the IETF, formed in 1986. Starting in 1969, technical decisions regarding the Internet were vetted by the user community using the RFC process established by Steve Crocker and the mediation powers of Jon Postel. Through the RFC process, any individual or group had a voice in Internet governance. Anyone could propose a modification and have it vetted by the IETF. This process has been formalized by a series of RFCs and is standard operating procedure for the ISOC, IAB, and IETF today.

This freewheeling approach to management should not work, but it does—perhaps because all successful RFCs become *Best Current Practices* (BCP) that the global owners and operators follow. RFCs are not mandatory, but rather prescribed. If they catch on and become integrated into the operation of the Internet, then they are elevated to BCP. Indeed, the open culture of the IETF has permeated the entire Internet culture and has had a profound impact on the way the communications sector has evolved. Excerpts from the RFC3233 below underscore two key features of Internet governance:

1. Its freewheeling—almost anarchical structure of governance, and
2. The culture of the Internet has evolved over 30 years.

According to RFC3233,

> [BCP9], the primary document that describes the Internet standards process, never defines the IETF. As described in BCP11 ("The Organizations Involved in the IETF Standards Process") [BCP11], the Internet Engineering Task Force (IETF) is an open global community of network designers, operators, vendors, and researchers producing technical specifications for the evolution of the Internet architecture and the smooth operation of the Internet. It is important to note that the IETF is not a corporation: it is an unincorporated, freestanding organization. The IETF is partially supported by the Internet Society (ISOC). ISOC is an international non-profit organization incorporated in the US with thousands of individual and corporate members throughout the world who pay membership fees to join. The Internet Society provides many services to the IETF, including insurance and some financial and logistical support. As described in BCP11, Internet standardization is an organized activity of the ISOC, with the ISOC Board of Trustees being responsible for ratifying the procedures and rules of the Internet standards process. However, the IETF is not a formal subset of ISOC; for example, one does not have to join ISOC to be a member of the IETF. There is no board of directors for the IETF, no formally signed bylaws, no treasurer, and so on.[20]

[19]IETF, Internet Engineering Task Force; IRTF, Internet Research Task Force.

[20]http://www.faqs.org/rfcs/rfc3233.html

As the number and scope of topics handled by the IETF broadened, the Internet Engineering Steering Group (IESG) was established by RFC3710 to manage the expanded number of working groups:

> The Internet Engineering Steering Group (IESG) is the group responsible for the direct operation of the IETF and for ensuring the quality of work produced by the IETF. The IESG charters and terminates working groups, selects their chairs, monitors their progress and coordinates efforts between them. The IESG performs technical review and approval of working group documents and candidates for the IETF standards track, and reviews other candidates for publication in the RFC series. It also administers IETF logistics, including operation of the Internet-Draft document series and the IETF meeting event.[21]

Most decisions that deeply affect the technical evolution of the Internet come from the IETF, which are ratified by the ISOC and implemented by vendors. It is a remarkably decentralized and unfettered system that has reinforced the freewheeling culture of individuals, groups, and corporations that collectively comprise Internet governance. Most of the political and international governance of the Internet come from the Internet Corporation for Assigned Names and Numbers (ICANN)—the governing body set up by the US government when the Internet was spun out of the NSF in 1998.

In 2015 the US Congress approved their separation from ICANN, signaling the acceptance of greater participation by non-US countries. Established in 1998 to manage the root DNS servers and handle the bookkeeping required to register URLs and IP addresses, the ICANN was set free from the US Department of Commerce in 2016. From that point on, ICANN became a multi-stakeholder entity with representation from around the globe.

6.7.2 ICANN Wars

The relatively self-governed Internet community does not always run itself without acrimony. In fact, there has been an abundance of disagreement over how the Internet should evolve, especially after it was outsourced by the US government. Most conspicuously was the so-called ICANN Wars, which raged for years after the commercial Internet was born in 1998.

In June 1998 the US National Telecommunications and Information Administration (NTIA) published the *White Paper* (Management of Internet Names and Addresses) in response to public comment on the *Green Paper*—an RFC-like proposal on how to commercialize the Internet. The NTIA proposed the formation of a nonprofit corporation—ICANN—which subsequently assumed responsibility for management of the DNS, allocation of IP address space, specification of protocols, and management of the root server system. ICANN does not register domain names itself. Instead, it delegates that responsibility to national registrars.

Nineteen directors who are broadly representative of the Internet community govern ICANN. Most members are appointed by their supporting organizations, but some are elected by members at large. For example, in 2003, the members of the ICANN Board were:

Internet pioneer Vinton Cerf (Chair)
Mexican academic Alejandro Pisanty (Vice-Chair)
European lawyer Amadeu Abril i Abril
California lawyer Karl Auerbach
Brazilian businessman Dr. Ivan Moura
US businessman Lyman Chapin
Canadian lawyer Jonathan Cohen
Mouhamet Diop
Japanese businessman Masanobu Katoh
Netherlands businessman Hans Kraaijenbrink
Korean academic Dr. Sang-Hyon Kyong
Dr. M. Stuart Lynn (ICANN President and CEO)
German journalist Andy Mueller-Maguhn
Japanese academic Dr. Jun Murai
Dr. Nii Narku Quaynor
German businessman Helmut Schink
Francisco A. Jesus Silva
US academic Dr. Linda Wilson

ICANN was envisioned to be more than an "FCC of the Internet," but fall short of "owning the Internet." But exactly what was the scope of ICANN's powers? In fact, a number of independent groups had other ideas about ICANN's power over the Internet. This difference of opinion evoked the ICANN Wars.

Dan Schiller—author of *Digital Capitalism*—called ICANN an "unelected parliament of the Web."[22] Karl Auerbach—ICANN board member in 2003—complained that ICANN was "essentially an organ of the trademark lobby."[23] Others accused ICANN of establishing policies that negatively impacted free expression and favored commercial interests over personal privacy. Milton Mueller lamented that the net's "role as a site of radical business and technology innovation, and its status as a revolutionary force that disrupts existing social and regulatory regimes, is coming to an

[21]http://www.faqs.org/rfcs/rfc3710.html

[22]Schiller, Dan, "Digital Capitalism Networking the Global Market System", MIT Press, Cambridge, (2000), 320 pp. ISBN 0-262-19417-1.

[23]http://www.icannwatch.org/

end."[24] Criticism was not restricted completely to individuals. Network Solutions, Inc.—the company that received a 5-year contract (1993–1998) to perform ICANN-like services on a temporary basis until the Internet was commercialized—complained in testimony to Congress that ICANN was out to destroy its business.

By the time you read this, ICANN may have changed again or been replaced. More than 200 leaders from government and business attended the Global Forum on Internet Governance, held in 2004 by the United Nations Information and Communication Technologies (ICT) Task Force. The purpose of this meeting was "to contribute to worldwide consultations to prepare the ground to a future Working Group on Internet Governance to be established by Secretary-General Kofi Annan, which is to report to the second phase of the World Summit on the Information Society."[25] The United Nations—like so many other Industrial Age organizations—was slow to understand the significance of the Internet. But once they "got it," they began to organize their own brand of governance. The United Nations created the Internet Governance Forum (IGF) in 2006 to continue the work of the World Summit on the Information Society (WSIS). The IGF brings together stakeholders from government, industry, and civil society to discuss Internet governance issues at its annual meetings.

6.7.3 ISOC

In 1992 soon after Congress directed NSF to commercialize NSFNet, Cerf and Kahn formed the ISOC, which has evolved into an umbrella organization, embracing social as well as technical issues.[26] Some topics of concern to ISOC are:

- Censorship
- Copyright
- Digital divide
- DNS
- E-commerce
- Encryption
- Privacy
- Public policy
- Security
- Societal
- Spam

[24]Mueller, Milton, "Ruling the Root: Internet Governance and the Taming of Cyberspace," MIT Press, Cambridge, (2004), 328 pp ISBN 0-262-63298-5

[25]http://www.circleid.com/channel/index/C0_1_1/

[26]www.isoc.org/

6.7.4 W3C

The startling success of the WWW and commercialization of the Internet prompted Berners-Lee and Al Vezza to form W3C to create WWW technology and standards in 1994. According to the W3C, its charter is to formally nurture the Web as the Internet has traditionally been nurtured by volunteers. The Internet is a highway; the Web is a transportation system:

> The Web is an application built on top of the Internet and, as such, has inherited its fundamental design principles.
>
> (1) *Interoperability*: Specifications for the Web's languages and protocols must be compatible with one another and allow (any) hardware and software used to access the Web to work together.
> (2) *Evolution*: The Web must be able to accommodate future technologies. Design principles such as simplicity, modularity, and extensibility will increase the chances that the Web will work with emerging technologies such as mobile Web devices and digital television, as well as others to come.
> (3) *Decentralization*: Decentralization is without a doubt the newest principle and most difficult to apply. To allow the Web to "scale" to worldwide proportions while resisting errors and breakdowns, the architecture (like the Internet) must limit or eliminate dependencies on central registries.[27]

The W3C has more profound objectives than making sure the Web is healthy. It seeks to take the Web to its next level. The WWW and its underlying HTML provided a standard *syntax* for information, but it did not define the *semantics* of the information. A "sentence" in HTML could be syntactically correct, but meaningless. For example, the English sentence "The four sides of a square are circles" is syntactically correct, but meaningless. So, Berners-Lee set about to add meaning to the WWW. In 1996, W3C began working on XML and the "semantic network."

XML consists of three major parts: a language for encoding information—both as a document and as a message; XSL (eXtensible Style Language) software for rendering the information on a display (browser, printer); and DTD (Data Type Definition), a language for specifying the meaning of the information. Think of XML as a language (English, French, Italian), DTD as a dictionary, and XSL as an interpreter. Whenever an XML message is received, the receiving computer looks into a corresponding DTD to find the meaning of the tags in the message and then uses XSL to render the message on the user's screen. This is like an English-speaking person using an English-to-French dictionary to parse and understand French.

Today all browsers support XML. In fact, XML is the technology used to solve many homeland security problems such

[27]www.w3.org/

as interoperability between different computer systems, sharing of information among people with different levels of security, and data mining to extract meaning out of databases.

By 1998 the Internet had matured to the point where it could be privatized. The NTIA (within the US Department of Commerce) produced a "Green Paper" describing how the Internet should be governed, how to transition the DNS to private ownership, proposed adding more global top-level domains (gTLDs) (such as .tv for television), proposing that trademarks be honored as Internet names, reducing the $50 DNS registration fee to $20, and setting aside 30% of the revenues from DNS registration for the Intellectual Infrastructure Fund (IIF). The Green Paper formalized Jon Postel's operation and created ICANN to sell blocks of names to several authorized resellers.

6.8 INTERNATIONALIZATION

The name and number assignment responsibility of ICANN was regionalized circa 2005 (RFC7020) subsequent to the involvement of the United Nations and concurrent with the emergence of the multi-stakeholder ethos of ICANN. Five regional Internet registry nonprofits govern five regions of the world:

- The African Network Information Center (AFRINIC) serves Africa.
- The American Registry for Internet Numbers (ARIN) serves Antarctica, Canada, parts of the Caribbean, and the United States.
- The Asia-Pacific Network Information Centre (APNIC) serves East Asia, Oceania, South Asia, and Southeast Asia.
- The Latin America and Caribbean Network Information Centre (LACNIC) serves most of the Caribbean and all of Latin America.
- The Réseaux IP Européens Network Coordination Centre (RIPE NCC) serves Europe, Central Asia, Russia, and West Asia.

These RIRs hold seats in ICANN and participate in an RFC process for technical innovation and in a United Nations-like governance process. These disparate organizations are tied together via memoranda of agreement and generally agree to protect unallocated IP numbers, continue to support the RFC process of bottom-up policy development, and act as the focal point of Internet community input. For example, in 2018, French President Emmanuel Macron made a high-level declaration of common principles for regulating the Internet and fighting back against cyber attacks, hate speech, and other cyber threats.

6.9 REGULATION AND BALKANIZATION

The technical structure of the Internet is a global network containing highly concentrated hubs that are critical to the operation of the entire sector. But understanding the technical structure of the Internet and WWW may be a small challenge compared with understanding the organizational and regulatory challenge posed by this vast infrastructure. The question of Internet ownership remains complex at the time of this writing. It does not belong to anyone or any company. Rather, it operates through a convoluted social network of volunteers, nonprofit organizations, government agencies, and for-profit corporations. It is a global social system—not controlled or augmented by any single government.

On the other hand, the competitive exclusion principle is evolving Internet ownership toward a handful of corporations. This process is governed by Gause's law, which suggests that it is inevitable that Internet ownership will eventually fall into the hands of one or a small oligopoly of corporations. This direction is contrary to the open and unincorporated culture of the Internet and its volunteers. Will there eventually be a clash?

The Internet has been called the "information superhighway," but it is radically different than the federally funded Interstate Highway System. It does not receive subsidies, nor is it considered a natural monopoly, even though the entire US society depends on the Internet as much as it does the Interstate Highway System. Destruction of the Internet would have severe consequences on the national economy. And yet there is no police force, fire department, or security force responsible for the Internet's safety or security.

The Internet has been called the most significant advance in human communication during the past 500 years.[28] And yet the FCC does not regulate it like radio or television, nor is it managed like roads, bridges, or power grids. Analog spectrum for radio and telephone broadcast (and cell phones too) is sold to the highest bidder for billions of dollars by governments around the world. According to the FCC, the electronic spectrum belongs to the interstate public and thus is subject to federal oversight. The Internet, on the other hand, is not restricted by any federal agency. Anyone can buy broadcasting rights for $20/year. ICANN and its authorized resellers literally give away one of the most valuable rights in human history—the right to broadcast to everyone in the world without a license.

But the age of the freewheeling Internet is over. The controversial net neutrality debate is still going on as this is written. Tim Wu describes net neutrality as "internet service providers [that] treat all data on the Internet the same and not block, speed up or slow down traffic based on paid prioritization or other preferences."[29] The FCC adopted net neutrality rules in 201, but reversed them in 2017. FCC commissioner Ajit Pai scraped the ruling on the basis that the FCC had no legal basis for regulating the Internet.

Net neutrality is a landmark ruling regardless of which way it ends up, because it establishes the notion that the Internet is something the US government can regulate. Net

[28]One can easily argue that the printing press was the most significant advance in communication prior to the Internet.

[29]https://www.law.columbia.edu/news/2017/11/net-neutrality-Tim-Wu-FCC

neutrality was the first Internet regulation by the United States. In 2018 the state of California declared net neutrality a state regulation of the Internet, but retreated into a wait-and-see status until questions of the FCC's right to regulate the Internet is answered by the courts.

A second landmark decision in 2018 by the EU court established the GDPR to regulate e-commerce companies operating in the EU. This profound legislation places bounds on Internet companies and business models that collect, aggregate, and sell consumer privacy and security. The GDPR is described in more detail in subsequent chapters (Hacking Social Networks).

The Internet has been compared with a global publishing and printing machine and a global vending machine. It provides merchants a global distribution channel that will soon reach all of humanity—for minimal cost. This has enormous consequences for e-commerce, societal change, and functioning nations. So far, no major government has imposed taxation on the Internet, and only minimal restrictions have been placed on spam, freedom of speech, and pornography. Will the United Nations react to this unprecedented freedom of expression? Will the Internet be banned in major parts of the world? And if it is, what does that mean to modern societies that are increasingly dependent on the Internet Age?

Like many other technological advances before it, the Internet and WWW have been exploited for both good and evil. The WWW supports human networks consisting of both terrorists and pen pals. It has been a vehicle for positive social change as well as social unrest. The Internet is destined to have a major impact on critical infrastructure sectors ranging from *lifeline* sectors (water, food, communications, and energy/power) to higher-level sectors such as public health and emergency services—and all the sectors in between.

The power of the Internet to change society and exploit information for good or evil purposes has led many governments to reign in its exposure to the global network. The editorial board of the *New York Times* predicts an eventual balkanization of the Internet. In the near future, the global Internet may splinter into three or four parts. The Internet could become four Internets—splinters run by China, Europe, the United States, and remaining free world and dominant companies such as Google.com. According to the *New York Times*,

> There's a world of difference between the European Union's General Data Protection Regulation, known commonly as GDPR, and China's technologically enforced censorship regime, often dubbed "the Great Firewall." But all three spheres—Europe, America and China—are generating sets of rules, regulations and norms that are beginning to rub up against one another. What's more, the actual physical location of data has increasingly become separated by region, with data confined to data centers inside the borders of countries with data localization laws.
>
> The information superhighway cracks apart more easily when so much of it depends on privately owned infrastructure. An error at Amazon Web Services created losses of service across the web in 2017; a storm disrupting a data center in Northern Virginia created similar failures in 2012. These were unintentional blackouts; the corporate custodians of the Internet have it within their power to do far more. Of course, nobody wants to turn off the Internet completely—that wouldn't make anyone money. But when a single company with huge market share chooses to comply with a law—or more worryingly, a mere suggestion from the authorities—a large chunk of the Internet ends up falling in line. The power of a handful of platforms and services combined with the dismal state of international cooperation across the world pushes us closer and closer to a splintered internet.[30]

The Internet is quickly becoming the most fundamental critical infrastructure—as critical as food, water, and power—because digital convergence is merging all communications together with sociopolitical as well as security threats. TCP/IP is the fundamental monoculture underlying these other infrastructure sectors. So whether the infrastructure sector is water, power, energy, emergency services, public health, agriculture, defense industrial base, critical manufacturing, or key resources such as nuclear power plants and government buildings, the Internet has emerged as the most vital component. This heavy reliance on TCP/IP makes the Internet the most critical of all critical infrastructures. Its trend toward control by irresponsible governments makes TCP/IP dangerous.

6.10 EXERCISES

1. What is the Internet?
   a. Any digital network
   b. Any packet switching network
   c. Any TCP/IP network
   d. Any Ethernet network
   e. All of the above
2. What is an Internet protocol?
   a. Rule for communication between networked devices
   b. IEEE 802.11 standard
   c. ARPANet predecessor to the Internet
   d. A Microsoft product
   e. Rules proposed by Al Gore, Vice President of the United States
3. In terms of DNS structure, the Internet is shaped like a:
   a. Hierarchical tree
   b. Mesh or grid graph
   c. Random graph
   d. Hamiltonian graph
   e. Complete or full graph
4. Packet switching networks were studied and invented by:
   a. Kleinrock
   b. Baran

[30]There May Soon Be Three Internets: America's Won't Necessarily Be the Best. https://www.nytimes.com/2018/10/15/opinion/internet-google-china-balkanization.html

c. Davies
d. All of the above
e. None of these people

5. Which of the following is an example of an Internet gTLD?
   a. .com
   b. name@earthlink.com
   c. An email attachment
   d. An email format
   e. www.CHDS.us
6. What does a DNS server do?
   a. Registers usernames
   b. Runs the Internet
   c. Translates a URL into an IP address
   d. Rebuilds the Internet
   e. Implements TCP/IP
7. Who invented and sent the first email?
   a. Ray Bradbury
   b. Ray Tomlinson
   c. Larry Roberts
   d. Ray Robinson
   e. Jon Postel
8. Which of the following is TRUE?
   a. Originally, TCP/IP was TCP.
   b. The ISO/OSI dictates what protocol is used by the Internet.
   c. TCP/IP is the same as the ISO/OSI Transport Layer.
   d. All Internet routers and switches use HTML.
   e. UNIX is the operating system of the Internet.
9. How much did the US government spend on the Internet during the period 1986–1995?
   a. $200 million
   b. $1 billion
   c. $1.5 billion
   d. $5 billion
   e. Nothing (it was commercialized by then)
10. Which protocol guarantees delivery of packets over the Internet?
    a. UDP
    b. TCP
    c. IP
    d. DNS
    e. SNMP
11. Which of the following government agencies commercialized the Internet?
    a. NTIA
    b. ICANN
    c. IANA
    d. IETF
    e. ISOC
12. Which one of the following is TRUE?
    a. Internet governance is top down, from the ISOC to the IETF.
    b. Internet governance is up to the US government.
    c. Internet governance is mainly through international volunteer organizations.
    d. Internet is owned by the IT-ISAC.
    e. Internet operation is regulated by the FCC.
13. Which one of the following is TRUE?
    a. The WWW and Internet are the same thing.
    b. The WWW is software.
    c. W3C and ISOC have overlapping powers.
    d. XML is an extension of HTML.
    e. None of the above are true.
14. Three routes exist between the sender and receiver of an email message in Figure 6.2. What happens if parts of one email message are sent along one route and another part is sent along a second route?
    a. The entire email message is retransmitted.
    b. The switches and routers use OSPF to correctly route the pieces.
    c. TCP flags the error.
    d. IP flags the error.
    e. DNS translates myname@myserver.com into 131.200.13.4.
15. Write an essay on how the Internet compares with:
    a. The Interstate Highway System
    b. Broadcast networks like radio and TV
    c. Mail-order catalog commerce
    d. Electric power utilities and the four interconnection grids of the United States

6.11 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. Most critical infrastructure is planned, while the Internet emerged without central control or a plan. Compare the results obtained by the planned Interstate Highway System versus the ARPANet in terms of planning, governance, and results.

B. Do your own research and examine what the pioneers such as Vint Cerf think about the security of TCP/IP and the Internet in general. Hindsight is 20–20. What do they say in hindsight?

C. It is relatively easy for a hacker to rewire the Internet by changing tables in routers. What do you think can be done to prevent unauthorized rewiring?

D. Explain in your own words what a monoculture is and why the Internet is considered a monoculture.

E. The Internet was designed to be robust against nuclear attack. Packet switching was part of the solution, but the Internet has self-organized. What is the implication of self-organization and Gause's competitive exclusion principle on the Internet's ability to survive a nuclear war?

7

CYBER THREATS

A *cyber threat* is a computer or computer network hazard. It is a potential attack that preys on weaknesses or flaws in hardware and software systems. An *exploit* is defined as an unauthorized action performed on an information system such as a corporate network, desktop personal computer (PC), enterprise server, Web site, factory control systems, SCADA network, or home computer. A *zero-day exploit* is a previously unknown or unrecognized exploit. A *remote exploit* is an unauthorized access to an information system from a distance—from across a network.

There are a number of highly varied types of threats ranging from malicious software designed to penetrate entire systems to *phishing* email exploits designed to betray users into giving out personal information and numerous other types of threats designed to cause a nuisance or very serious theft of intellectual property, financial gain, and espionage. Perhaps the worst exploit is a *rootkit*, because it yields complete control of a computer to a hacker and requires an entire rebuild of the victim computer to remove. We categorize all of these as malware (**mal**icious soft**ware**) to simplify the terminology. As of 2019, the following types of malware roamed the wilds of the Internet:

- Adware or spam: unwanted advertisements.
- Trojan: malware that appears to be a valid application or part of the computer's operating system, but is actually malicious. A Trojan typically has the same name as the operating system module it replaces.
- Bot: malware for distributed denial-of-service (DDoS) attacks evolved into a network of agents under the control of a botherder.
- Ransomware: malware that encrypts your files for the purpose of extortion. The decryption key is offered for sale, but often the keys are not delivered.
- Rootkit: a Trojan that gets past the security checkpoints of your operating system lies in wait for a command from the criminal to take control of the victim computer.
- Spyware: malware that scrapes keyboards (keylogger), screens, and video camera feeds for the purposes of invasion of privacy or exfiltration of passwords and personal information.
- Virus: a self-replicating malware that is activated by an authorized owner or operator of a computer system.
- Worm: a self-replicating malware that spreads and activates on its own.

This chapter is about exploits—the potential unauthorized acts against the *information technology (IT) sector* for the purpose of gaining control, stealing information, destroying data, and denying service to the authorized users of IT systems. We assume that the information systems of greatest interest are nodes connected to one another via the Internet. The links connecting these nodes are any TCP/IP connection, whether it is a wired or wireless communication link. However, because of digital convergence, the Internet connects not only Web-based systems but also non-Web systems

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

such as factory control, energy and power grids, and transportation systems. The IT sector provides an infrastructure for almost every other CIKR system—and every one of them is vulnerable to malware attacks.

We must assume a highly percolated Internet as described in Chapter 6. This heightens the threat, because, as illustrated earlier, percolation is a form of self-organized criticality that magnifies the consequences of normal accidents. From the previous chapter we learned that the Internet is highly structured around a hub-and-spoke architecture containing super-spreaders. Because of these super-spreaders, we must assume that malicious software exploits will reach all parts of the Internet with very little expense, time, or effort on the part of the perpetrator. This assumption carries over to interdependent systems that use the Internet such as energy pipeline systems, the power grid, banking system, transportation systems, and municipal water systems. This interdependency makes the study of cyber threats of the highest importance.

In this chapter, the following concepts are explained through a combination of theory and real-world example:

- The threat surface contains many vulnerabilities: concern about computer security has been on a steady rise since the 1970s, but only since commercialization of the Internet since 1998 has concern gone from mild to critical. Like a battlefield, the attack surface of the Internet is actually several layers of attack surfaces. Vulnerabilities differ from layer to layer. Unfortunately, these layers are target-rich surfaces.
- Cyber thieves are generally divided into several major groups: script kiddies are inexperienced novices seeking notoriety; black-hats are knowledgeable hackers seeking fame and fortune, and crackers are even more knowledgeable criminals with more serious damage in mind—typically international thieves and militants. A relatively new kind of attacker—the nation-state hacker—joined the list of malicious perpetrators when Stuxnet was launched to take out the Iranian nuclear program. Attacks are euphemistically called exploits, because cybercriminals and nation-states attempt to exploit weaknesses in information systems.
- The tools of the exploit trade are viruses (malicious self-replicating and user-activating programs), worms (malicious self-replicating and self-activating programs that spread via the network), phishing to defraud users via email or fake Web pages, Trojans that supplant portions of an operating system with malicious code, and, more recently, weaponized bots (offensive worms like Stuxnet). These tools are used to render a Web site unusable by denying access (denial of service [DoS]), infect files and databases, inflict loss of information on the operation, destroy industrial control systems, and stop or lower worker productivity. In some cases, an exploit can result in the hacker remotely taking control of a target computer for financial gain or to cause physical damage.
- Break-ins typically begin with a minor infraction such as an unauthorized login and escalate to more serious infractions using backdoor programs (malicious programs stored on the victim's computer), Trojan horses (deceptive programs that look innocent but are actually malicious), and botnet zombies (other people's computers that are used to launch an attack on even more computers). Weaponized worms and viruses are more focused—they attempt to damage or take control of a specific target, such as a power plant, water treatment plant, or uranium purification facility.
- Viruses are an older technology that typically infects disks, thumb drives, and application software. Worms are the more likely type of exploit. They spread like an epidemic throughout the IT sector by exploiting TCP/IP flaws, unattended ports, weaknesses in operating systems, email weaknesses, and miscellaneous flaws in software at all levels. Bots are malicious programs that inhabit other people's IT systems (zombies) and lie in wait for signals from a botherder. Bots connect with one another and therefore create a network on top of the Internet.
- The highly connected IT sector is an extreme example of a cascade network, which means it has the same vulnerabilities as self-organized networks studied earlier: worms spread like epidemics in human populations. Of particular concern is the very high spectral radius of the Internet, which greatly magnifies the spread of malicious software. The most effective countermeasure is to harden the most connected hubs in the autonomous system (AS-level) network. The spread of online worms can be virtually stopped by hardening 2–3% of all AS-level servers.
- AI countermeasures: Defenders are applying automated detection software based on machine learning that looks inside of other software to compare and detect malware based on its structure. When structures match known worms or viruses, the machine learning software blocks or deletes the malware.

7.1 THREAT SURFACE

A century ago the US national economy depended on railroads and heavy industries to create wealth. Today, the US economy is heavily dependent on information, and information is captured, stored, moved, processed, and delivered by information systems. These information systems have replaced many of the Industrial Age physical systems with a far more fragile virtual system. The train-robber

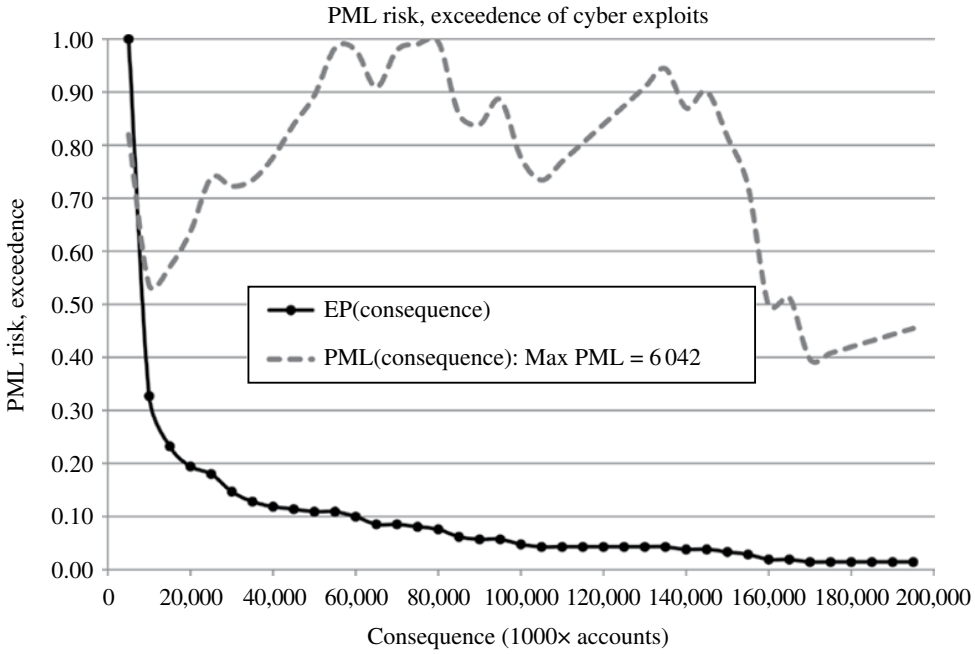


FIGURE 7.1 The exceedence probability and PML risk profile of a sample of 211 computer security exploits of all sizes suggests that computer security is a high-risk challenge due to the long tail of the exceedence probability and upswing in the PML curve as consequence increases.

has been replaced by the cyber thief and cyber fraud—the so-called script kiddies and black-hats that prey on vulnerabilities in information systems.[1] Curiosity motivates script kiddies who use automated tools that are readily available over the Web to probe other people's computer systems. More pernicious are the black-hats—people that are driven by more serious motivations. Hackers typically break in because they can, while crackers break in to destroy or steal information. Both are knowledgeable experts that often develop their own malicious programs—mainly worms.[2] Cyber threats are not acts of nature, but instead are manufactured by hackers and crackers.

Since the release of Stuxnet, nation-states have become active as crackers of financial, governmental, and military IT systems. News broadcasts report of theft, propaganda, misinformation campaigns, and threats against CIKR on a daily basis. Cybersecurity has become a major industry in both commercial and military organizations throughout the world. The so-called threat surface is very active (see Fig. 7.2).

Computer security is a major problem both the United States and countries worldwide. Since 1998 the amount of intrusions has escalated and made it hard for consumers to trust e-commerce sites such as banks and popular Web sites known as watering holes—highly popular sites that attract criminals due to their high traffic. A social network site, for example, is a target for pedophiles, and a law enforcement site is an attractive target for hackers trying to extract information on criminal cases, information on law enforcement members or private sector partners. Doxing police officers or firefighters is the practice of releasing personal information about first responders, such as their home address, phone numbers, or family members.

The following sample of exploits carried out in 2018–2019 illustrates the diversity of threats and targets. Figure 7.1 shows the exceedence probability curve and PML risk curve for 211 exploits listed prior to 2018. Note that computer security tends to be a high-risk challenge, because PML risk rises steadily as consequences increase:

- Marriott hotels: Hackers accessed the reservation database and stole the guest information of 500 million customers, including phone numbers, email addresses, passport numbers, reservation dates, and some payment card numbers and expiration dates.
- Hackers accessed India's government ID database and stole 1.1 billion citizens' identity and biometric information.
- Ouora: Criminals obtained 100 million customer accounts including names, email addresses, encrypted passwords, data from user accounts, and users' public questions and answers.
- MyHeritage: Black-hats obtained 92 million email addresses and encrypted passwords of users who signed up for the service.

[1]Script kiddies are amateurs out for fun and glory. Black-hats are serious professionals working for financial gain.

[2]A worm is a self-activating program that spreads through a computer network like an epidemic.

- Google+ profiles: Hackers obtained 52.5 million customer accounts including name, employer and job title, email address, birth date, age, and relationship status. Google shut down Google+ in 2019.
- Saks and Lord & Taylor: A hacking group called JokerStash stole more than 5 million credit and debit cards and advertised their sale to other criminals.
- T-Mobile: Hackers obtained 2 million accounts that included encrypted passwords, personal data, account numbers, billing information, and email addresses.
- British Airways: Hackers compromised 380,000 accounts from the airline's bookings made on the Web site and app, obtaining the credit card payment information.
- Chip manufacturers: Two families of flaws in microprocessor chips called Meltdown and Spectre allow access to higher-privileged parts of a computer's memory. Hackers may exploit them to steal data from deep within memory even while applications are running securely on a machine. This vulnerability has been verified on Intel chips and AMD and ARM processors and is not easily patched because it requires changes to hardware architecture.
- Facebook: The myPersonality app mishandled Facebook user data by sharing "information with researchers as well as companies with only limited protections in place."

Government agencies at both the state and federal level report a high number of consequential cyber intrusions:

- State of Texas: 3.5 million affected in April 2011, including theft of Social Security numbers, dates of birth, and driver's license numbers.
- South Carolina Department of Revenue: 3.6 million Social Security numbers and 387,000 taxpayers' credit and debit card numbers stolen.
- Tricare: Exposure of 4.9 million military hospital and clinic patient accounts containing personal data, including full names, home addresses, phone numbers, and Social Security numbers.
- Georgia Secretary of State Office: 6.2 million voter information records including Social Security numbers were stolen.
- Office of the Texas Attorney General: 6.5 million records containing sensitive information and Social Security numbers stolen from a voter database.
- Virginia Department of Health Professions: 8.3 million.
- US Office of Personnel Management (OPM): 21.5 million.
- US Department of Veteran Affairs: 26.5 million.

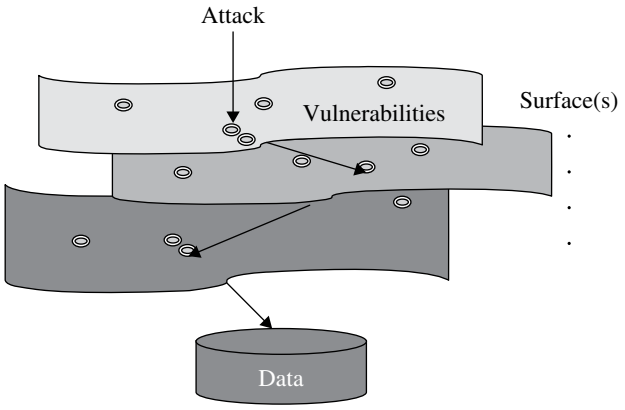


FIGURE 7.2 The concept of an attack surface or surfaces where computer security struggles take place parallels the attack surface of a battlefield. Most systems are layered to block access along a kill chain shown here as a path to data.

- National Archives and Records Administration (NARA): 76 million.
- US Voter Database: 191 million.
- Singapore: Names and addresses in the Singapore government's health database and some patients' history of dispensed medicines. Information on the prime minister of Singapore was specifically targeted.

The foregoing examples describe a battlefield-like struggle between e-commerce sites and government databases versus attacking armies reminiscent of an attack surface describing the land, air, and sea campaigns in war. The description fits and has led many practitioners to describe computer security as a struggle between good and evil over an attack surface (see Fig. 7.2).

The cyber attack surface is actually many layers thick, depending on the information system and its architecture. One surface may be the simple login and password layer of a cloud computer where various applications run. Another layer may be the application itself, and yet another layer may be the underlying operating system and storage subsystem. Each layer may contain security constraints and vulnerabilities.

Keeping with the military metaphor, a path from layer to layer through vulnerability holes as shown in Figure 7.2 establishes a kill chain. If completed, the hacker eventually gains access to data, control, or both. The job of a defender, then, is to block the attacker somewhere along the kill chain. This is the kill chain approach to computer security. It is one of several frameworks for defeating adversarial hackers. Chapter 10 describes the kill chain framework in more detail.

7.1.1 Script Kiddies

Adrian Lamo, "the homeless hacker," cracked computer systems at *The New York Times*, Yahoo, Bank of America,

Citigroup, and Microsoft using free computers at places like coffee shops and libraries. He found flaws in his victim's information systems, exploited them, and then told the companies about their vulnerabilities. Lamo may have pioneered a racket used a decade later—cyber extortion of World Wide Web companies by promising to not attack their sites in exchange for money. Lamo was eventually caught and ordered to pay approximately $65,000 in restitution and was sentenced to 6 months of home confinement plus 2 years of probation.

In a strange turn of events, Lamo exposed Pfc. Bradley Manning—the soldier who released classified information in an infamous exploit that became known of as WikiLeaks. Lamo reported Manning's betrayal after chatting with him for 6 days in May 2010. He eventually notified the FBI in Sacramento, California, which led to the capture and sentencing of Manning to 35 years in prison. President Obama eventually commuted her sentence.

Lamo died in March 2018 for unknown reasons, although he had a history of seizures. He was only 37 years old. Some fellow hackers accused Lamo of betraying the hacker ethos. But Lamo said, "Had I done nothing, I would always have been left wondering whether the hundreds of thousands of documents that had been leaked to unknown third parties would end up costing lives, either directly or indirectly."[3]

Dark Dante, whose real name is Kevin Poulsen, worked for *SRI International* by day and hacked by night. His most famous exploit won him a brand new Porsche automobile. Each week the Los Angeles radio station KIIS-FM awarded a $50,000 Porsche 944 to the 102nd caller following a preannounced sequence of songs. When the song sequence triggered the calling frenzy, Poulsen took over the station's phone system, blocked out all other callers, made call number 102, and drove away with the prize.

More seriously, he hacked into an FBI database containing wiretap information, perhaps to punish the FBI. Law enforcement dubbed him "the Hannibal Lecter of computer crime." When he was captured, authorities found so many hacking devices they compared him to James Bond. Poulsen was captured in a supermarket after a 17-month pursuit and served 51 months in jail and fined $56,000. Poulsen later became a senior editor for *Wired News*, specializing in cybercrimes and cybercriminals. His most prominent article exposed 744 sex offenders who exploited *MySpace.com* profiles.

Script kiddies are accomplished amateurs who hack systems for many nonfinancial or nonviolent reasons. Often they are motivated by the challenge or bragging rights. They also acquire malware from the Web rather than build it themselves, although the previous examples show a keen intellect.

[3]https://www.theguardian.com/us-news/2018/mar/16/adrian-lamo-dead-chelsea-manning-wikileaks

7.1.2 Black-Hats

Perhaps the best-known black-hat is Kevin Mitnick, the self-proclaimed "hacker poster boy." The Department of Justice described him as "the most wanted computer criminal in United States history." Mitnick was the subject of two movies—*Freedom Downtime* and *Takedown*. Mitnick began his career as a small-time thief, hacking the Los Angeles bus-ticketing system for free rides. He then dabbled in phone phreaking—hacking the telephone system to make free long-distance calls. His online bio says, "[My] hobby as an adolescent consisted of studying methods, tactics, and strategies used to circumvent computer security."[4]

Mitnick was eventually caught and convicted of stealing software. He served 5 years, of which about 8 months was spent in solitary confinement. He became a computer security consultant, author, and speaker, appearing on television shows: *60 Minutes*, *The Learning Channel*, *Court TV*, *Good Morning America*, *CNN*, and *National Public Radio*. He is the author of two books: *The Art of Deception* (2002) and *The Art of Intrusion* (2005).

7.1.3 Weaponized Exploits

According to some experts, a cyber "Pearl Harbor" is unlikely, because such an operation would be highly complex, require extreme coordination effort, and result in dubious damage [1]. A more likely scenario is that future black-hats will use cyber attacks asymmetrically—as a force multiplier in concert with a physical attack. For example, a cyber attack might be used to interrupt emergency services, manipulate traffic control signals, hinder disaster recovery, and so forth, in concert with a bomb, biological, chemical, or other physical assault.

But a cyber Pearl Harbor might be feasible using a weaponized exploit as demonstrated by the *Stuxnet*—a recombinant virus designed specifically to target the uranium centrifuge facility in Iran. Stuxnet was not the work of script kiddies or black-hats, but rather the offensive work of nations. It was the first widely known weaponized virus used by a country, although others have preceded it. Stuxnet is interesting because of two features—it was designed as an offensive weapon, and it is recombinant—made from several other software fragments. The recombinant nature of Stuxnet is perhaps its most serious property, because it means malicious software of the future will mutate into more powerful and sophisticated threats.

Cybersecurity experts believe Stuxnet was launched circa June 2009 via a USB memory stick that was inserted into a Windows PC. A previously known flaw in Internet Explorer was used to penetrate the Window's operating system. Additionally, Stuxnet used stolen security certificates to get

[4]http://mitnicksecurity.com/media/Kevin_Mitnick_Bio_BW.pdf

past computer security at Iran's uranium refinery. It targeted the Siemens' industrial control system being used to control the centrifuges at Natanz, Iran. By giving commands to speed up, the exploit was able to destabilize and destroy the centrifuges.

Stuxnet is an example of a *recombinant virus*—it was 10× more complex than previous viruses. It combined a 2008 Explorer virus with a virus known as *Zlob*, plus a 2009 print virus, plus a Siemens' Step7 password exploit. A typical computer virus is 15,000 lines of code. Stuxnet exceeded 500,000 lines.

Governments are becoming a bigger threat to the Internet than script kiddies and black-hats. In May 2007, President Bush authorized the National Security Agency (NSA) to attack the cellular phones and computers operated by insurgents in Iraq. The insurgents were posting videos of roadside strikes on the Internet to recruit followers [2]. The NSA operators hacked into the insurgent's network and set a trap sprung by waiting US soldiers.

A dispute between the Russian Federation and neighboring Georgia over a region called South Ossetia located on the Russian–Georgian border initiated an effective cyber exploit designed to augment physical conflict between the two countries' military forces in August 2008. It began when Georgian forces launched a surprise attack against South Ossetian separatist forces on August 7. Russia responded the next day by sending troops into Georgian territory. Cyber attacks were launched against Georgian governmental Web sites prior to the physical attacks. The exploits primarily defaced public Web sites and denied service to a number of other sites.

7.1.4 Ransomware and the NSA

According to the *New York Times*, "Millions of people saw their computers shut down by ransomware, with demands for payments in digital currency to have their access restored. Tens of thousands of employees at Mondelez International, the maker of Oreo cookies, had their data completely wiped. FedEx reported that an attack on a European subsidiary had halted deliveries and cost $300 million. Hospitals in Pennsylvania, Britain and Indonesia had to turn away patients. The attacks disrupted production at a car plant in France, an oil company in Brazil and a chocolate factory in Tasmania, among thousands of enterprises affected worldwide" [3].

The ransomware malware used in these attacks was derived from the Equation Group collection of hacking tools developed by the Tailored Access Operations (TAO) group within NSA. The Shadow Brokers claimed responsibility for the exfiltration of NSA's Equation Group tools for breaking into other people's computers. The WannaCry ransomware is derived from one of these tools as are dozens of other highly effective and potent computer hacking tools (see Fig. 7.3).

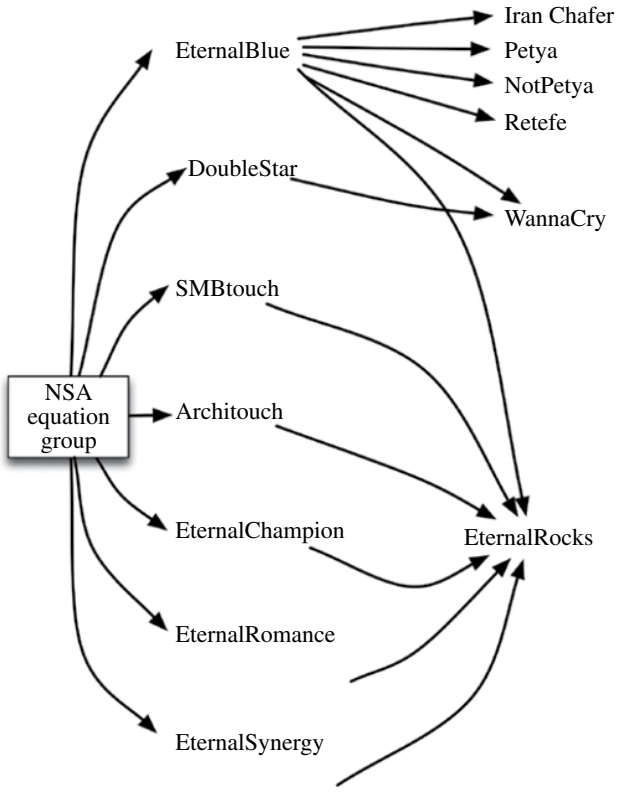


FIGURE 7.3 This "family tree" shows the heritage of some of the malware derived from leaked computer cracking tools called Equation Group, developed by the NSA. The Equation Group refers to a collection of tools, while the Shadow Brokers refers to the individual or organization that leaked them.

The Shadow Brokers' leakage of the Equation Group tools continues to have a far-reaching impact. Exploits against firewalls and SNMP (Simple Network Management Protocol) control of network equipment from Cisco and Juniper continue to occur in the wild. Another example is EternalRocks, also known as Doomsday. "It is one of the first instances of a ransomware program that uses stealth. It also called itself WannaCry to hide from security researchers. Once a computer is infected by it, it stays hidden on the host computer, secretly installs Tor Browser, and then makes a connection to its servers. After 24 hours, the server will begin to self-replicate the malware. This ransomware does not seem to have a kill switch yet, unlike some of the WannaCry variants. So far, EternalRocks just seems to infect computers, however it has been warned that this worm could be weaponized at any time. As this ransomware has stealth capabilities, it is unknown how many computers are infected with it at the moment."[5]

EternalRocks is an example of a remote access Trojan (RAT). It creates a backdoor for remote access to machines it infects. It is also an example of recombination in malware.

[5]http://malware.wikia.com/wiki/EternalRocks

Hackers compose more capable RATs by combining other malware into super-malware. Stuxnet was one of the first weaponized RATs to use this idea. In Figure 7.3 we see that EternalRocks recombines many other types of malware to increase its potential for working its way past firewalls and other countermeasures to deliver a payload without being detected or blocked.

RAT researchers have reported upwards of 450,000 infected computers by one or more of the Equation Group tools and derivatives. Fortunately, computer antivirus software is an effective countermeasure. Unfortunately, IoT devices, point-of-sale terminals used in retail, and ATM cash dispensers are not as well protected. These systems must be constantly updated to fend off recombinant mutations of Equation Group tools.

Fundamentally, malware preys on operating system and application program flaws. As software complexity rises, the opportunity for flaws also rises. Complexity equals opportunity. This is not restricted to software, however. The Meltdown and Spectre family of flaws in hardware opened up microprocessors to exploitation of speculative execution of machine-level instructions.[6] Nearly every modern microprocessor performs speculative execution to speed processing. Instead of executing instruction in order, speculative execution performs instructions as soon as hardware is available to perform operations such as add, subtract, copy, and multiply, regardless of order. This works out, most of the time, but when it does not, the hardware must back up and recalculate the operations in proper order. In order to back up and recalculate, the hardware must save results of previous computations. The results of these speculative operations are saved in protected memory. Unfortunately, speculative execution was found to lay bare privileged information in protected memory—a security breach that is open to exfiltration.

7.2 BASIC VULNERABILITIES

The most common exploits prey on flaws in software programs and hardware. Due to the complexity and size of most software in use today, many flaws or "holes" exist in the operating systems, application programs, and hardware that they run on. Software flaws—called *defects*—are often discovered years after consumers have deployed the software. Once a defect is discovered, the software manufacturer may offer a repair—a *patch*—that fixes the problem. Unfortunately, many of these patches are never installed, leaving the information system vulnerable. One of the most effective countermeasures to combat cyber attack—the patch—is often overlooked.

[6]https://arstechnica.com/information-technology/2018/11/intel-cpus-fall-to-new-hyperthreading-exploit-that-pilfers-crypto-keys/

Hackers and crackers use a variety of methods for penetrating corporate and home systems. One of the oldest is called *war dialing*, where a hacker programs his or her computer to dial all the telephone numbers listed in a telephone book, until a modem is sensed. Once the war dialing computer senses a modem tone, it repeatedly sends login and password combinations—words taken from the English dictionary. If the password is a proper English word, the war dialer will eventually discover it. The same idea is used on email addresses—randomly trying out all combinations until one is found without a secure password.

War dialing is a tedious brute-force method of breaking into someone's computer, but since it is done by another computer, it is easy and inexpensive for the hacker. Today the large number of open wireless access points encourages a variation of war dialing called *war driving*. A mobile hacker equipped with a laptop computer simply drives around a neighborhood until an 802.11 Wi-Fi signal is detected and then begins exhaustively sending a series of username and password codes (generated from a dictionary) until the login is successful.

Once in, a hacker or cracker will attempt to escalate his or her access privileges. Can the invader open up password files to get more usernames and passwords? Can he or she intercept other user's email? Is the victim's address book unprotected? Is the corporate database accessible from the login?

In some cases, the professional cyber thief can store a program on the cracked system for use at a later time. This is called a *backdoor*—a program that the hacker activates from outside of the security zone of the cracked system. A backdoor program may lie dormant for a long period of time before it is activated or activate itself periodically. It can look like an authorized part of the system but instead become destructive. A *Trojan horse* program is a deception—it looks valid, but it is not.

War dialing and war driving are not the only means of hacking into a system. A large number of exploits come from employees or trusted associates—the so-called *insiders*. But perhaps the most disturbing exploits come from outside the organization. If information systems can be attacked from anywhere in the world, no infrastructure sector is secure. Anyone in the world can attack any Internet-connected system located anywhere else in the world with inexpensive equipment and knowledge of how computers and networks operate. In fact, the construction of malicious programs for the purpose of carrying out *remote exploits* has become a cottage industry of virus, worm, and Trojan horse software developers.

Attacks from inside the organization and its information system perimeter are called *insider attacks*. Whether an attack comes from the inside or outside, exploits are not difficult to initiate. The tools already exist, and for the most part, they can be acquired at little expense. Many of these

tools are available from the Web itself. They fall into the following general categories:

- Virus programs (user-activating software that spreads via files, etc.).
- Backdoor programs (black-hat takes remote control).
- Trojan horse programs (deceptive software).
- Worm programs (self-activating software that spreads via a network).

These tools are used for a variety of nefarious activities, including, but not limited to:

- Stealing passwords or credit card information.
- Taking control of a remote computer or network.
- Destroying or corrupting files and databases.
- Using a remote computer to spread viruses and worms to others.
- Turning a remote computer into a *zombie*—a computer that launches a subsequent DoS attack on a Web site or corporate network.

A *virus* is a malicious self-replicating program. A *worm* is a malicious self-replicating program that spreads through a network. A *Trojan horse* is a data file or program containing a malicious program. It is a computer program that appears to be harmless but actually does damage. For our purposes, viruses, worms, and Trojan horses are all the same—*malicious software* used by hackers and crackers to do damage or take control of information systems. In the following description of exploits and how they work, we will treat viruses, worms, and Trojan horses as tools of the black-hat trade.

7.2.1 The First Exploit

In 1988, Robert Tappan Morris—a 23-year-old PhD student at Cornell University in Syracuse, NY—remotely launched the first cyber worm aimed at a MIT machine located miles away in Cambridge, MA. The worm quickly infected and disrupted 6000 computer systems and their users across the United States. In some cases, the worm forced users to disconnect from the Internet to stop the worm.

How did this happen? Morris had discovered two flaws in the operating system of the Internet's servers. These flaws allowed him to gain unauthorized access to machines all over the Internet. Once inside the target machine, the worm used the target machine's routing tables and usernames/passwords to find new victims. The worm copied itself onto other machines, where the process was repeated.

The worm attempted three exploits on target machines:

1. Execute a remote command that gives the attacker access to the target machine.
2. Force a so-called buffer overflow exploit on the target machine, which inadvertently relinquishes control to the attacker.
3. Access the email program on the target machine, and command it to download the virus onto the target machine.

Morris was caught, charged with computer fraud and abuse, and found guilty on May 4, 1990, in Syracuse, New York. He was sentenced to 3 years' probation, levied a $10,000 fine, and required to contribute 400 h of his time to performing community services. Estimates of financial loss range from $100,000 to $10,000,000. Today, Morris is an accomplished computer scientist working at a university.

The Morris exploit illustrates several features of cyber threats. First, flaws in the software of networked computers make it possible for hackers and crackers to gain access to remote machines. Cyber threats depend on these flaws to gain a foothold. Second, the malicious program replicates by copying itself onto other vulnerable machines. In other words, a computer virus or worm works much the same way that a biological virus or worm does—it reproduces itself and travels to a new host, where the process is repeated. Malicious programs can infect many remote systems at the speed of the Internet. Third, this historical example illustrates what is still true: hackers may cause millions of dollars of loss, but they typically get modest sentences.

Viruses existed long before worms. In the early days of the PC, viruses traveled by infecting floppy disks, document files, and application programs. They did not depend on the Internet, but rather, they spread through physical contact. One of the oldest exploits used special tracks on software distribution disks, called the *boot record*. When the infected disk is inserted into a PC, the boot record is copied into the main memory of the PC. Once inside, the infected boot record made copies of itself on every disk inserted into the PC. The virus spread to new target machines whenever a human computer user shared the disk with another user. Today a computer user inadvertently activates viruses, whereas a worm spreads on its own.

Other viruses work through other vectors. A virus might attach itself to a document file, such as an Excel or Word file. Wherever the file goes, the virus goes also. The user activates the virus when it is loaded into his or her computer. Trojan horse viruses frequently traveled this way before the Internet became widely used.

Microsoft Office products are designed to allow a programmer to embed a program inside of a Word or Excel document. These programs are called *macros*.[7] When activated, macros perform routine tasks or add additional capability to the Office application. But macros are vulnerabilities in Microsoft products, because a macro can be a Trojan horse. A

[7]Macros are written in Visual Basic and activated by the user whenever the document is loaded into an Office application.

hacker can embed a malicious macro in a Word document and sent it as an attachment to millions of PCs. When the user opens the attachment, the macro activates and does its damage.

Worms can be thought of as mobile viruses, because they copy themselves onto target computers but do not require a user to initiate them. They are a favorite of black-hats because they can infect the entire Internet with little time, effort, or expense on the part of the attacker. Hence worms pose an asymmetric threat to the Internet, SCADA networks, financial networks, power grids, and telecommunication networks.

Worms can do anything any other software can do. Worms have been known to email the entire contents of a victim's hard disk to others, install a backdoor Trojan horse on the victim's computer, observe and record a user's keyboard keystrokes (keylogger), launch DoS attacks, disable antivirus software, and steal or destroy a victim's files. Worm exploits were the most frequent type of exploit in 2004, but have declined in use since 2005.

There are five fundamental ways that worms propagate from computer to computer:

1. Fundamental flaws in TCP/IP.
2. Unprotected or open input/output ports on target machines.
3. Operating system flaws: buffer overflow exploits.
4. Email protocols and attachments.
5. Flawed applications and system software.

7.2.2 TCP/IP Flaws

In the previous chapter we surveyed TCP/IP's historical past and observed that it was designed over 30 years ago to enable the United States to regain leadership in missile technology, save money by sharing expensive computers among university and research labs, and withstand thermonuclear attack from the former Soviet Union. But it was not designed to withstand cyber attacks. Unfortunately, there are many known flaws in TCP/IP that make it extremely easy to exploit.

Each TCP/IP packet contains both source and destination address. The source address identifies the server that sent the packet, and the destination address identifies the intended recipient of the packet. These IP addresses are clear—they can be read and changed by anyone clever enough to intercept and modify them. IP addresses are obtained from the distributed DNS server system described in Chapter 6. The path from one user to another is established by a collection of routing tables stored in switches and routers. An attack on the DNS servers and routing tables is a severe attack on the Internet. How can these be hacked?

TCP/IP's address-in-the-clear vulnerability stalled the largest banking network in the country in 2003. SQL Slammer was launched on the weekend of January 24–26, 2003. It caused systems running Microsoft SQL Server to generate a massive number of messages with random source and destination addresses. This generated a flood of traffic between pairs of Microsoft SQL Servers, crowding out all other traffic on the Internet. One of the infected computers happened to also be connected to the Bank of America ATM network. When this computer stalled, it also stalled the Bank of America ATM network.

The Bank of America ATM network was affected because the bank's financial network and Internet access were both hosted on the same machine. The worm got past this machine's firewall, and it became flooded with millions of short messages. Normally, this machine exchanged information between ATM and non-ATM networks, but on Monday morning, the server was so loaded down with messages generated by the SQL Slammer that it was useless. All 13,000 Bank of America ATMs became unusable until port 1434 (used by Slammer) was filtered.

Another flaw in TCP/IP is responsible for another type of DoS attack. The SYN flooding exploit works because TCP/IP was designed to be simple—not necessarily flawless. SYN Flooding is possible because of the three-way handshake used by TCP to establish a connection between two computers A and B. Figure 7.4 shows what is supposed to happen and what can happen when SYN flooding is used to overload a server with ceaseless unresolved SYN messages.

In Figure 7.4, system A (sender) initiates a connection to system B (receiver) by sending a SYN message.[8] System B responds to the SYN request by returning a SYN followed by an ACK within a reasonable time interval. Sending a confirming ACK from sender to receiver confirms the three-way handshake. Once the ACK is received, system A sends the message to system B.

But what if the three-way handshake never completes? An exploitation of this initiation protocol occurs when system A *never* returns the expected ACK corresponding with its initial SYN. System A (sender) and system B (receiver) shake hands by exchanging a SYN and ACK as before. But the receiver never gets an ACK. Instead, system B (receiver) gets a stream of more SYNs. This ceaseless stream keeps the receiver busy doing nothing but waiting for ACKs that never arrive. Meanwhile, system B stores the pending SYN requests until its memory overflows, causing system B to collapse.

SYN flooding is an elementary DoS exploit. If millions of SYNs are sent to a single receiver, both network and receiving system get bogged down with handshaking, which leaves little time to process valid messages. The hacker can magnify the number of sending systems by hijacking zombie computers (computers taken over by the hacker without the owner's knowledge). Millions of zombies can be infected

[8]SYN is short for synchronize and ACK is short for acknowledge—two hold over signals from the days of teletypes and Western Union "email" delivery.

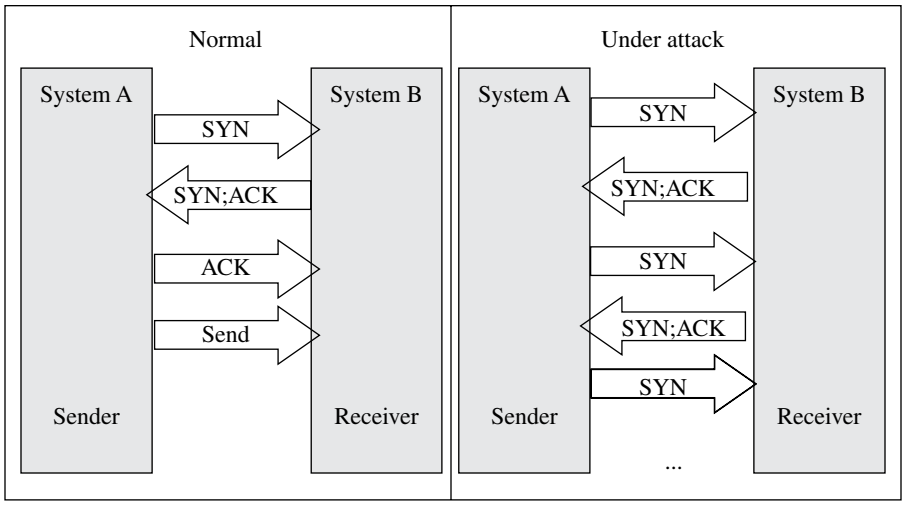


FIGURE 7.4 TCP/IP is intrinsically flawed because of its simplicity.

with malicious software timed to all send their SYN flooding messages to the target. The remedy is to close the port or filter the stream of SYNs coming from the zombies.

DoS attacks using TCP/IP flaws are commonplace, although they are not as frequent as other exploits. But they can be dramatic. *Code Red* was launched against the Whitehouse Web site on October 21, 2002. Within hours it was detected on millions of computers around the globe. Here is how Code Red worked:

- The worm enters the target computer through its port 80.
- It finds and infects the Microsoft Internet Information server software.
- It copies itself onto other targets generated at random for 20 days.
- Then it goes dormant until a certain date, when all copies are activated.
- Millions of distributed copies flood the Whitehouse server with messages.

Code Red inundated the www.whitehouse.gov server with messages generated by zombie computers selected at random as the worm spread. The number of zombies numbered in the millions, because the worm replicated itself for 20 days before launching the DoS exploit against www.whitehouse.gov. It infected servers on all continents including Europe, Africa, Russia, China, and the North and South America.

Fortunately, Code Red used the numerical IP address of the Whitehouse server, instead of the symbolic www.whitehouse.gov. The DoS attack was diverted by simply changing the DNS address book, pointing www.whitehouse.gov to a server with a different IP address. The Whitehouse administrators essentially *spoofed* the attacker.[9]

[9]Spoofing means that source address of packets returned from DNS are changed to something else, which changes the identity of the sender.

Code Red illustrated several vulnerabilities in the Web:

1. Port 80 (the port used by all Web browsers) was used, showing that a worm can travel through commonly used ports.
2. Code Red showed that a simple worm could cause widespread damage.
3. DoS attacks are simple but effective.

DoS attacks do not destroy information or cause physical damage. They simply render the Web site they attack useless. In emergency or national security crises, information and IT systems are essential to the operation of police, fire, military, medical, power, energy, transportation, and logistical systems. Without IT systems, modern information societies are crippled, if not permanently damaged.

7.2.3 Open Ports

Code Red used port 80 to travel through cyberspace. Every computer has ports—doors through which information enters and leaves a system. Ports are numbered from 1 to 65,535, but only a few are actually used in a given computer. Some well-known ports are:

Port no.	How it is used
25	TELNET
80	HTTP
443	HTTPS
21	FTP
110	POP3
25	SMTP
1433	SQL Server
53	DNS

For example, port 21 is the preferred doorway to a commonly used data transfer program called FTP (File Transport

Protocol). FTP provides a fast way to transfer large files. It also provides a fast way for worms to spread through exploitation of ports. Recombinant viruses called *Sasser.C* and *Sasser.D* swept through Windows XP and Windows 2000 systems in 2004 using FTP—mostly infecting home computers (500,000 to 1,000,000). The *Sasser* worm is not a single worm, but a series or strain of worms that have mutated over time much like a biological virus mutates as it adapts to threats.

The worm scans random IP addresses for exploitable systems. When one is found, the worm exploits the vulnerable system through a *buffer overflow* exploit. Here is how a typical Sasser worm works:

- The worm initiates an infection by creating a remote program via port 9995.
- The remote program creates an FTP program on the target computer, which downloads the remainder of the malicious program onto the target computer, thus completing the infection.
- Now the infected target accepts any FTP traffic on port 5554, which gives the attacker access to the target computer.

Note that Sasser uses a combination of open ports and buffer overflow. What is buffer overflow?

7.2.4 Buffer Overflow Exploits

One of the oldest and still most difficult exploits is called a *buffer overflow* exploit. Essentially, this exploit uses that fact that a computer does not know the difference between data and program code. All information looks the same to a computer, but if data is interpreted as code, then the infected computer can be fooled into accepting malicious code as if it were data. In a buffer overflow exploit, a virus, disguised as data, is sent from the attacker to the victim, but once it arrives at the target computer, it turns into a malicious program! How is this possible?

Figure 7.5a shows what is supposed to happen when data enters a computer operating system from an open port. Normally, the operating system acts as an intermediary between the outside world and the application (user) program. Input data is temporarily stored in a storage area called a *buffer*, along with a return address that tells the operating system where to return control once the data has been transferred. After the buffer fills up, the return address is used to return control back to the user program. The user program then transfers the input data into its own processing area.

Figure 7.5b shows how a buffer overflow can be exploited to wrest control away from the operating system (and the user program) and turn control over to a malicious program. Data enters the target computer as before, but this time it overflows the storage buffer. In fact, it writes over the return address stack and inserts a new return address that returns control to the buffer and stack itself. The operating system uses the hacked return address to pass control to the malicious program, which now resides in the storage buffer or stack. What was thought to be data actually turns out to be a malicious program.

Perpetrators of buffer overflow attacks must discover the size of the storage buffer and return stack of each system by trial and error. That is, they have to guess where to place the malicious return address and viral code. This is done by launching thousands of buffer overflow attacks containing one, two, three, … hundreds of different trial return addresses, until one works.

In July 2003 the *Win32:Blaster* worm (aka *msblast.exe*) used port 135 and a buffer in Windows to spread throughout the Web. Here is how the buffer overflow exploit worked:

- Exploits buffer overflow in the Microsoft Windows Remote Procedure Call (RPC) interface.
- Scans 20 hosts at a time, trying to connect to port 135.
- When an open port is found, the worm copies itself to the host using TFTP.
- Activated whenever Windows is started (via Windows registry).
- Can force Windows to restart.

RPC is the remote procedure call mechanism that allows two computers to communicate with one another. TFTP is trivial FTP, and the Windows registry is a table inside of Windows that holds the names of programs that are allowed to run on a user's computer. The Windows registry is one of the primary targets of hackers because it contains access rights to everything in the computer. Registry attacks are commonplace, and new zero-day exploits aimed at breaking into the registry occur every day.

7.2.5 DDoS Attacks

Open ports, buffer overflows, and various flaws in software provide contamination vectors for the spread of cyber viruses and worms. These vectors can be exploited in thousands of computers at once to turn innocent victims into collaborators in DoS attacks. When harnessed together, these *zombies* create a *DDOS* attack. The DDoS is one of the most effective exploits known, but fortunately, they do less harm than an exploit that erases files or steals documents.

A DDoS exploit starts by infecting a large number of zombies with an idle virus that lies in wait until a certain date arrives or signal occurs. At some specified later time, the zombies simultaneously flood a single target computer with meaningless data. The objective is to overload the target with

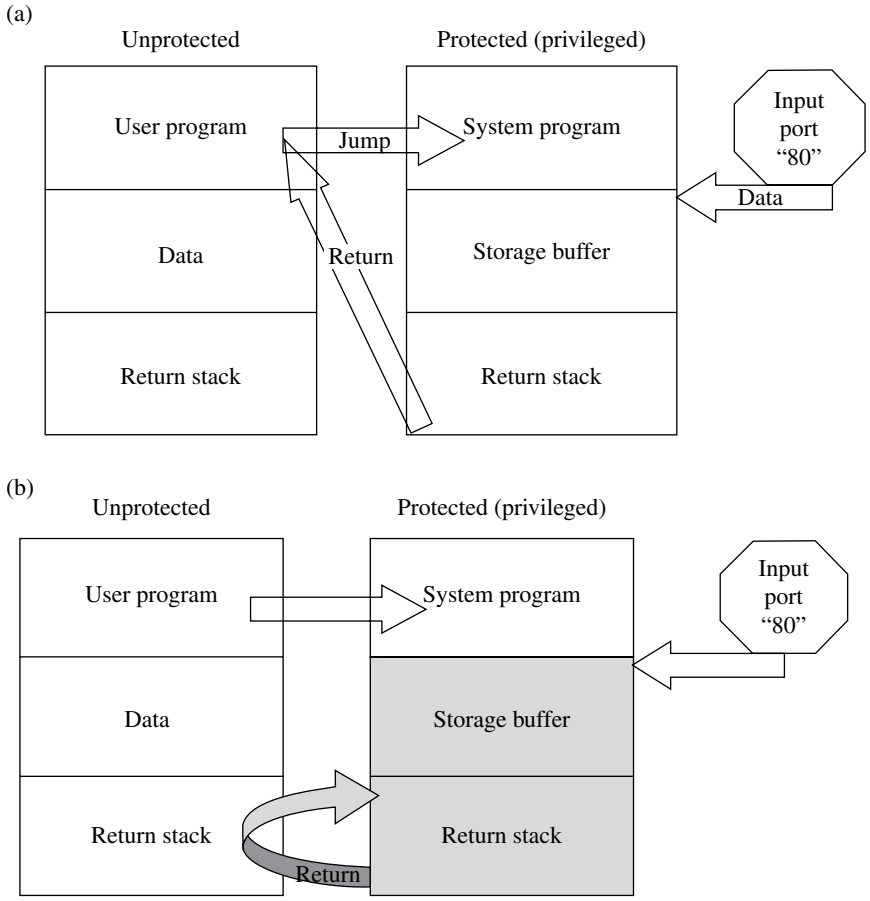


FIGURE 7.5 Buffer overflow exploits enter a computer as data but overwrites portions of code to fool the computer system into treating the data as executable code. (a) Information is read into a storage buffer as data where it is passed on to the user. (b) Stored information changes the return address, pointing the computer to the previously stored information that is now interpreted as code.

messages, rendering it useless for ordinary processing. Figure 7.6 shows a diagram of the two-phase DDoS exploit.

A 15-year-old Canadian teenager calling himself and his exploit *MafiaBoy* launched a DDoS strike against the most popular e-commerce sites in February 2000. This worm flooded Amazon.com, Buy.com, CNN.com, eBay.com, E-Trade.com, Yahoo.com, and ZDNet.com with millions of messages, resulting in an estimated loss of $1.7 billion in revenue. The MafiaBoy worm electronically recruited an army of zombie computers around the world, which in turn flooded the e-commerce servers with thousands of simultaneous requests for service, forcing them to shut down for several hours. The teenager was fined $250 and given an 8-month jail sentence.

The MafiaBoy exploit illustrates how DDoS attacks work, and it also illustrates another unfortunate fact: billions of dollars of damage can be done with very inexpensive software. It does not even take a clever person to launch a DDoS attack. Anyone can download the software and turn it loose in the wild. Even more disturbing are the social consequences of hacking: an underage offender can render billions of dollars of damage but suffer almost no punishment, or punishment that is extremely disproportionate to the amount of damage caused.

7.2.6 Email Exploits

Email exploits are carried out by hackers using malicious programs that predominately spread by email attachment. They depend on the victim activating the virus when it arrives as an attachment. The victim activates the virus by clicking on it or saving it to his or her local storage. They typically modify the Microsoft Windows registry, which gives the hacker unlimited access to the victim's computer.[10] Email viruses can do all the things that other viruses can such as installing backdoors, keyloggers, and compromising security and data integrity.

One of the most virulent email exploits in 2002 was w32.klez.e@mm, also known simply as the *Klez virus*. It spread

[10]The registry is where Windows keeps the names and authorizations for every program that is allowed to run on a Windows machine.

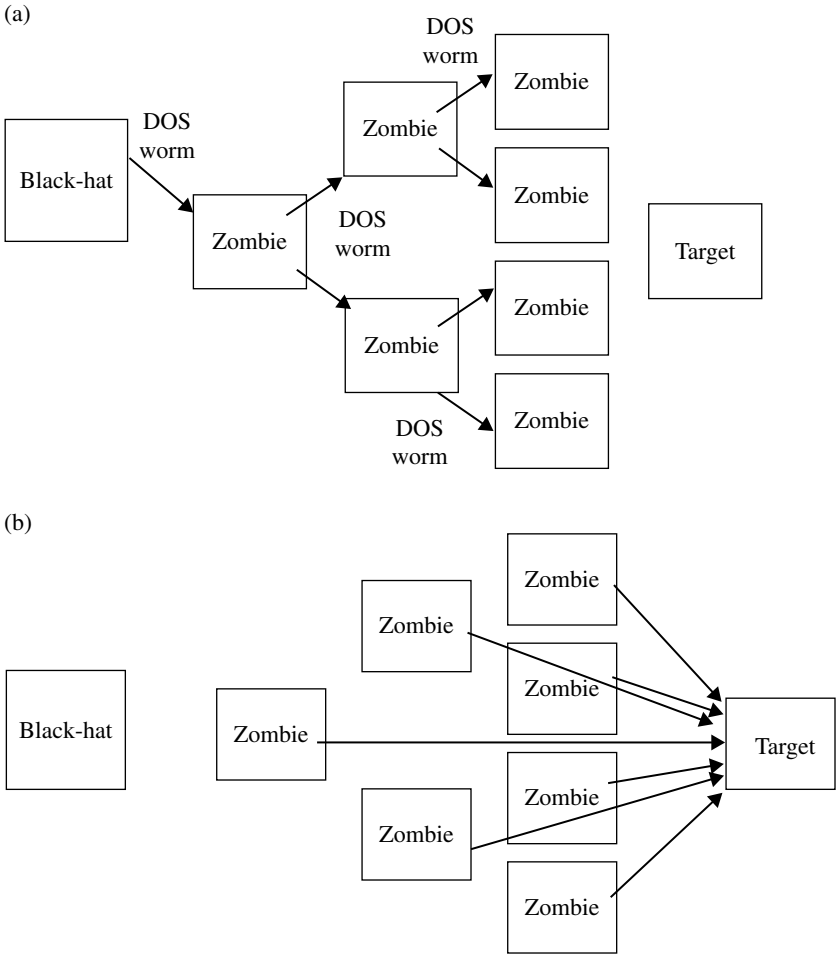


FIGURE 7.6 DDoS recruits innocent zombies to participate in a denial-of-service attack against a single target computer. (a) Phase 1: Spread the worm to thousands of zombies. (b) Phase 2: Zombies flood the target with messages.

by reading the entries on the victim's Microsoft Outlook address book. Klez tried to disable the user's antivirus programs, copy itself to network disks, and mail itself to all entries in the user's Outlook address book.

In 2003, *Bugbear* used email attachments and MIME attachment vulnerability to spread and install a backdoor, keylogger, and its own SMTP engine that sent spoofed email using the victim's address book. Bugbear could do serious damage to the target computer, including deletion of the user's files.

7.2.7 Flawed Application and System Software

The number of exploitable flaws in application and system software is legion. Malicious software enters a user's computer through flaws and weaknesses as follows:

- HTML and XML as clandestine message software.
- HTTP may leave open access doors.
- ActiveX: Code from the Web can access your computer.
- SMTP and POP3: Email can give away your password.
- SNMP: Manages networks, but also opens it to the outside.
- SOAP/XML: RPC can be used against you.

HTML and XML uses tags to tell a Web browser such as Microsoft Internet Explorer what each line of data means. But if Internet Explorer encounters an unknown tag, it simply skips over the line of data and continues looking for meaningful tags. What if the unknown tag is actually a malicious program? This is called *steganography* and has been used by spies for thousands of years to conceal messages within other messages.[11] It is also a tool of hackers.

[11]Steganography is the art and science of hiding information by embedding messages within other, seemingly harmless messages. It is used for both good and evil. Steganography is used to electronically protect intellectual property by embedding watermarks in digital documents, as well as to conceal secret messages.

HTTP (Hypertext Transport Protocol) is the protocol that dictates how a Web browser communicates with a Web server. There are several vulnerabilities in this fundamental software. First, it communicates in the clear—transmissions can be intercepted and substituted by unscrupulous people looking for credit card names and numbers. Second, version HTTP 1.1 leaves sessions open, because repeated opening and closing of sessions is inefficient. But open sessions can be used like open ports. Hackers can exploit port 80, which is the port used by http. Once port 80 is hacked, the currently running session can be hacked too.[12]

HTTPS (HTTP Secure server) and SSL (Secure Socket Layer) should be used instead of http, when security is important. HTTPS/SSL encrypts transmissions between server and Web browser. Most secure e-commerce applications—such as credit card buying over the Internet—are run on HTTPS/SSL encrypted sessions. Never enter a credit card number of banking account number into a Web site unless the URL starts with https://.

HTTPS/SSL transmissions may still be vulnerable to ActiveX programs that are transmitted between server and desktop. ActiveX is a Microsoft system for downloading programs and running them on the user's computer. Secure ActiveX programs ask for the user's permission and require a security certificate. But most users grant access to every ActiveX program without knowing what each program does! The ActiveX program may be a virus that destroys information or installs other malicious programs on the user's machine. How can a user know?

ActiveX programs should not be allowed to write to a user's local disk drive or alter a Windows registry file. Without prior knowledge of what the ActiveX program will do, granting access is like inviting a stranger to take over your house for the weekend! Web browsers can be set to block ActiveX downloads as well as data snippets called *cookies*, which may contain personal information.

ActiveX software has been employed by unscrupulous merchants and advertisers to promote their products and services. Called *spyware* for good reason, these ActiveX programs collect information about the user so that the unscrupulous merchant can target him or her for advertising, personalization, or privacy violations. For example, file-sharing music pirates used *Kazaa* in 2003 to download spyware to home computer users and then subsequently spammed the unsuspecting users with pop-up ads.

An EarthLink.com study found more than 29 million spyware-related files on 1 million of their subscriber's computers. Dell Computer customer support reported 12% of their support calls were complaints about spyware. In 2004, Microsoft attributed 50% of reported crashes of Windows XP to spyware.

[12]A session and an application is almost identical, so hacking an open session is tantamount to hacking a running application.

Professional black-hats possess even deeper knowledge of how computers and the Web operate. Because of this knowledge, they have devised complex exploits that go far beyond the scope of this book. Exploits involving the SNMP and POP3 email servers are known for exposing passwords; the network management protocol, SNMP, used to maintain the hardware of the Internet is also vulnerable to hacks; and the SOAP/XML protocols used by e-commerce companies is vulnerable to knowledgeable hackers and crackers. The list continues to get longer, and the exploits continue to get more sophisticated.

7.2.8 Trojans, Worms, Viruses, and Keyloggers

Trojans, worms, viruses, and keyloggers are the most prominent and troublesome malware to date. Trojans are the most common form of malware and Trojans rage in sophistication from simple to complex. The following is a survey of common malware found in the wild, circa 2019. It is an ever expanding list:

- Recall that a rootkit provides continued privileged access to a computer while actively hiding its presence, because a rootkit reaches inside of the privileged execution level of a computer operating system. The word root refers to the protected Admin account on Unix and Linux systems, and the word kit refers to the malware that implements the exploit. Rootkits are generally associated with Trojans, worms, and viruses.
- A backdoor is defined as a type of Trojan that allows access to a system by bypassing its security and gaining access to systems undetected. Backdoors are installed and hidden so that a criminal can access a victim computer at any time. They assist spyware for gathering information on a device and sending it to the invading hacker.
- A new type of Trojan appeared in 2015 known as file-encrypting ransomware. A ransomware Trojan enters a system to encrypt data so the owner/operator cannot use it. Rather than exfiltrate the data, ransomware extorts owners by locking up files so the owner cannot use them. Typically, a bitcoin fee is paid to unlock the files. Bitcoin is untraceable.
- A particularly prolific information-stealing campaign throughout 2018, and into 2019, came in the form of the Emotet Trojan, which, among other things, steals data, monitors network traffic, spreads through networks, and drops other Trojans onto victim systems. For example, Emotet spreads TrickBot. Both Trojans are constantly updated with new capabilities such as the ability to steal passwords and browser histories.
- Anubis spread to 93 different countries targeting 377 financial apps to exfiltrate account details. It records

audio, sends SMS messages, makes calls, and alters external storage.

- LoJax gets its name from LoJack, an anti-theft product from developer Absolute Software. The rootkit is a modified version of a 2008 release of LoJack (then called Computrace). Its design ensured that even if a thief made major changes to a computer's hardware or software, the rootkit would remain intact. LoJax connects to servers believed to be operated by Fancy Bear, a hacking group that works under the direction of the Russian government.

7.2.9 Hacking the DNS

The hierarchical DNS is the heart of the Internet. Recall that its job is to convert URLs into IP addresses. Users communicate with the DNS through port 53, which is open to authorized and non-authorized users, both. If a flaw or malware attack on a DNS server succeeds in changing the correspondence between URL and IP address, the Internet is essentially rewired. Information can flow from anywhere to anywhere else at the whim of a criminal. Therefore, DNS security is extremely tight and the root DNS servers are kept in hidden locations.

Unfortunately, the DNS has flaws. In 2008, white-hat researcher Dan Kaminsky revealed a flaw known as the "DNS rebinding attack," which would have allowed hackers to take control of the DNS itself. Fortunately, Kaminsky kept his secret to himself and informed ICANN. The flaw was fixed.

But in 2018 ICANN issued a warning that the DNS had been under a "multifaceted attack" for many months. It seemed the state-sponsored black-hats were going after the Internet infrastructure itself.

> In January, security company FireEye revealed that hackers likely associated with Iran were hijacking DNS records on a massive scale, by rerouting users from a legitimate web address to a malicious server to steal passwords. This so-called "DNSpionage" campaign, dubbed by Cisco's Talos intelligence team, was targeting governments in Lebanon and the United Arab Emirates. Homeland Security's newly founded Cybersecurity Infrastructure Security Agency later warned that U.S. agencies were also under attack. In its first emergency order amid a government shutdown, the agency ordered federal agencies to take action against DNS tampering.[13]

Part of the problem was that local DNS operators were not using DNSSEC, a public–private key encryption protocol similar to certificate authorities. The hackers gained access through various phishing exploits that fooled operators into believing the state actors were authorized.

This highlights the vulnerability of the distributed nature of the DNS—access to servers in emerging nations may not be as protected as they should be. In 2018, only about 20% of the DNS servers were using DNSSEC. Poor policy and practice, not technology, is the vulnerability.

7.3 BOTNETS

In the early 1990s, clever IRC (Internet Relay Chat) programmers invented *IRC bots*—programs that simulated users of the IRC online community. These programs evolved into over-networks—networks on top of the Internet network—mostly for doing useful but tedious IT chores. These networks-within-networks turned black when nefarious individuals began using the technology to spoof unsuspecting consumers. The most common spoof became *spam*—unsolicited email typically sent as bulk email to email lists that have been collected from online chat rooms, Web sites, newsgroups, and viruses that harvest unsuspecting users' address books. Botnets are swarms of unwanted worms that collect and disseminate spam and invade IT installations for nefarious purposes.

Control of a botnet is under a *botherder*—someone or some organization that directs the botnet from a distance. For example, a single botherder might direct the botnet to remotely recruit zombies to spread the bot and do its dirty work. A typical botnet and botherder works as follows[14]:

- The botherder launches worms to infect millions of zombies with a bot.
- The bot collects user information via the zombies and reports the personal information back to the botherder.
- The botherder sells botnet services and information to a third-party spammer.
- The spammer gives spam messages to the botherder who in turn instructs the zombies to send the spam throughout the botnet's distribution channel.

In 2010 *Rustock* was the largest known botnet at the time—a collection of 1.6–2.4 million unsuspecting zombie computers herded by organized crime located in St. Petersburg, Russia.[15,16] Rustock used the popular IRC to link together massive numbers of zombies. IRC is a kind of *peer-to-peer* (P2P) network like the ones used by music and movie pirates for media sharing. P2P over-networks are cheap, powerful, and resilient, because they are parasitic—living off of established IT infrastructure. But, unlike the public-switched

[13]https://techcrunch.com/2019/02/23/icann-ongoing-attacks-dns/

[14]https://en.wikipedia.org/wiki/Botnet#Illegal_botnets

[15]http://www.MessageLabs.com, www.Honeynet.org

[16]http://en.wikipedia.org/wiki/Russian_Business_Network

Internet, P2P networks are distributed with relatively low spectral radius. Hence they are more resilient than the IT infrastructure they depend on.

Typical botnets can easily spew out 250 spams/min, night and day. During the month of August 2010, 1 in every 300 emails contained a virus, spam, or worm; 1 in 500 contained a phishing exploit; and over 4000 Web sites were being blocked by a botnet every day. In 2009, *Zeus*—the largest known botnet in the United States with 3.6 million zombies in tow—sold millions of usernames, passwords, account numbers, and credit card numbers using keylogger technology.

Due to their immense size and capacity to spam the globe, botnets pose a serious threat to the very existence of the Internet. For example, if the 2 million zombies in the Rustock botnet were to simultaneously emit high bandwidth spams, the load on the Internet could cripple or halt traffic all over the world. The botherder could extort companies, regions, and even nations that depend on Internet traffic for everyday business, commerce, and military coordination.

Estonia's experience in 2007 illustrates the power of botnets. Estonian Web sites were hammered for days after the government ordered the relocation of the Soviet-era war monument *Bronze Soldier* from the center of Tallinn to its suburbs. Ethnic Russians rioted for two days. DDOS attacks on government Web sites were so severe that many agencies were forced to discontinue service into and out of Estonia for several days.

Over 1300 people were arrested, 100 were injured, and 1 person was killed in the rioting. Some of the attacks were traced back to Russia, but eventually a 20-year-old Estonian student named Dmitri Galushkevich was arrested and charged with launching the DDoS from his PC. According to NATO, the cyber attack on Estonia did not qualify as a military attack. If it had, and if Russia or some other country had launched the attacks, other NATO countries would have been obliged to come to Estonia's rescue.

At the time this was written, it is not clear what the future of botnets holds for the IT sector. At one extreme, botnets can potentially take over the Internet and parasitically dominate its host. This would essentially turn control of the IT sector over to botherders. At the other extreme, botnets could be a temporary phenomenon, because policies and practices described later in this chapter can banish them, altogether. In short, it is up to nation-states to enact policies and laws banning botnets, if we want the global IT infrastructure to be safe and secure.

7.3.1 Hardware Flaws

In 2018 a new vulnerability appeared in computer hardware that remained unsolved without radical change to processor hardware itself. Meltdown and Spectre are actually two families of flaws based on a very subtle timing and concurrency vulnerability found in nearly all processor designs. They are both forms of memory leakage, and they are both derived from speculative execution found in nearly every modern processor.[17] Speculative execution speeds up processing performance by 15–20% through a "look ahead" algorithm embedded in the design of the processor. Blocking or removing it, reduces performance by 15–20%.

Meltdown exploits speculation within a single machine instruction, typically during a memory fetch. Information leakage may occur when memory is copied into the processor before checking privileges. Spectre exploits branch prediction and out-of-order execution of instructions. A simple example of a Spectre exploit is described here.

Consider a simple branch operation written in a high-level language, translated into low-level machine language, and executed speculatively:

$$\text{If}\left(\text{A} == 0\right) \quad \text{then} \quad \text{B} = 1 \quad \text{else} \quad \text{C} = 2$$

The machine code produced by this programming language statement is sequential. That is, the value of A is tested to see if it is zero, followed by a jump to B = 1 or C = 2. One of the following sequences is carried out at the machine level:

$$\text{A} == 0, \quad \text{so B is set to 1}$$
$$\text{A} \neq 0, \quad \text{so C is set to 2}$$

But what if both sequences are done at the same time, and one ignored, depending on the outcome of the test A == 0? Suppose B is set to one at the same time C is set to two while A is being tested. If A == 0, the value for C is discarded and 1 is assigned to B. Alternatively, if A ≠ 0, the value of B is discarded, and 2 is assigned to C. All three operations can be performed in one step instead of two. Performance is improved 100%, because all three operations were done at once.

This is called speculative execution because the hardware speculates on the outcome. The results of B = 1 and C = 2 are held in an internal cache memory deep within the processor. Only after A == 0 completes its test is one of the cached values assigned to B or C. Regardless of the outcome of the test, the value of both B and C is already computed. This saves a step. Speedup averages out to approximately 15–20% of total branch operations and varies in number of operations that can be computed in parallel.

So how does a criminal exploit speculative execution? If malware is running in another partition of the processor, it has access to the same cache as all other applications. A fast-acting malware program can snatch the temporary values of A, B, and C from the cache before speculative execution moves on to another branch in the program. The cache temporarily holds values that may be confidential but left

[17]https://en.wikipedia.org/wiki/Spectre_(security_vulnerability)

unprotected in cache memory. This represents an information leak in the design of the hardware.

Spectre is a family of exploits that snatch data from cache memory. Most of the time the data is mundane and of little significance. But what if it is an unencrypted password or key to an encryption algorithm? Cache memory does not have the same protection as protected memory, so data is (temporarily) left in the clear.

Software patches have been applied to operating systems to avoid Spectre exploits, but they reduce performance and in some cases they are ineffective against a cleverly designed exploit. As of 2019, there was no known solution to the Spectre vulnerability that did not reduce performance or require redesign of nearly all processors from nearly all manufacturers. It remains an unsolved problem for hardware designers.

Abu-Ghazeleh et al. note that "The Spectre and Meltdown vulnerabilities presented a conundrum to the computing industry because the vulnerability originates in hardware. In some cases the best we can do for existing systems—which make up the bulk of installed servers and PCs—is to try to rewrite software to attempt to limit the damage. But these solutions are ad hoc, incomplete, and often result in a big hit to computer performance. At the same time, researchers and CPU designers have started thinking about how to design future CPUs that keep speculation without compromising security" [4].

7.4 CYBER RISK ANALYSIS

Now that we are equipped with a fundamental understanding of cyber threats, it is possible to build a general model of cyber exploit risk against a generic computer system. This model may be used to reduce the risks facing single computer threat–asset pairs or risks facing a corporate data center's threat–asset pairs. It is not, however, a model of a networked system of computers. For Internet system risk, consider the top 2000 autonomous systems in the Internet circa 2005 analyzed in the next section.

Figure 7.7 contains a fault tree model of threat–asset pairs typically found in a computer system. The threat–asset pairs are:

Threat	Asset
Spoofing	TCP/IP
DDOS	
HTTPS/SSL	
Keylogger	Ports
Trojan horse	
Browser	Software flaws
ActiveX	
Media player	
Keylogger	Attachments
Trojan horse	
Address book	

Assuming all consequences are \$10,000, all threats are 100%, and all vulnerabilities are 10%, initially, what is the best use of investment dollars to reduce risk in this fault tree? Even though TV = 10% for every threat–asset pair, the overall vulnerability is 68.6%. Why is this so high? The answer lies in the OR logic of the fault tree. Any one or multiple threats may occur, which drives the threat of zero, one, two, three, or more exploits to 68.6%. Even though each individual threat is relatively small, the possibility of one or more threats is large.

The second observation addresses the allocation resources to reduce risk. Figure 7.8 shows the return on investment (ROI) curve for risk reduction investments. ROI drops below \$1.00/\$ when the total investment exceeds \$3000. At \$3,000, risk declines from \$11,000 to \$7,943, and vulnerability drops from 68.6% to 56.4%. An investment of twice this amount—\$6000—is required to lower vulnerability below 50%, but the ROI is much lower—\$0.76/\$. It is extremely difficult to reduce risk when assets are threatened by multiple hazards.

This example illustrates the challenge of cybersecurity. The large array of threats makes protection very difficult and expensive. In fact, most computer system operators have never performed a rigorous risk analysis of the information systems under their care. According to a 2003 study, 43% of system administrators surveyed did not know how their systems got infected.[18] Furthermore, they estimated that exploits had a minor financial consequence: 75% of the exploits detected caused less than \$100 damage, according to the IT managers. This may seem small, but cyber exploits affect millions of computers once they spread. Therefore, an exploit that infects a million computers really cost millions of dollars.

The 2003 survey also concluded that 30% of the exploits caused some loss of data—typically due to a virus. If we also assume each data loss incident cost \$100, then such exploits cost \$30 million/million victims. Reducing the risk of a single IT installation by \$1000 may not seem like much, but it is multiplied by potentially millions of similar IT installations, and the system risk can run into the millions or even billions.

7.5 CYBER INFRASTRUCTURE RISK

The foregoing fault tree analysis of threat–asset pairs suggests it is difficult to protect individual IT installations but the consequences are relatively small. However, when multiplied by the millions of IT installations connected to one another via the Internet, the accumulated consequences can become very large. The IT infrastructure is so heavily dependent on the Internet that it becomes necessary to return

[18]http://www.avast.com/

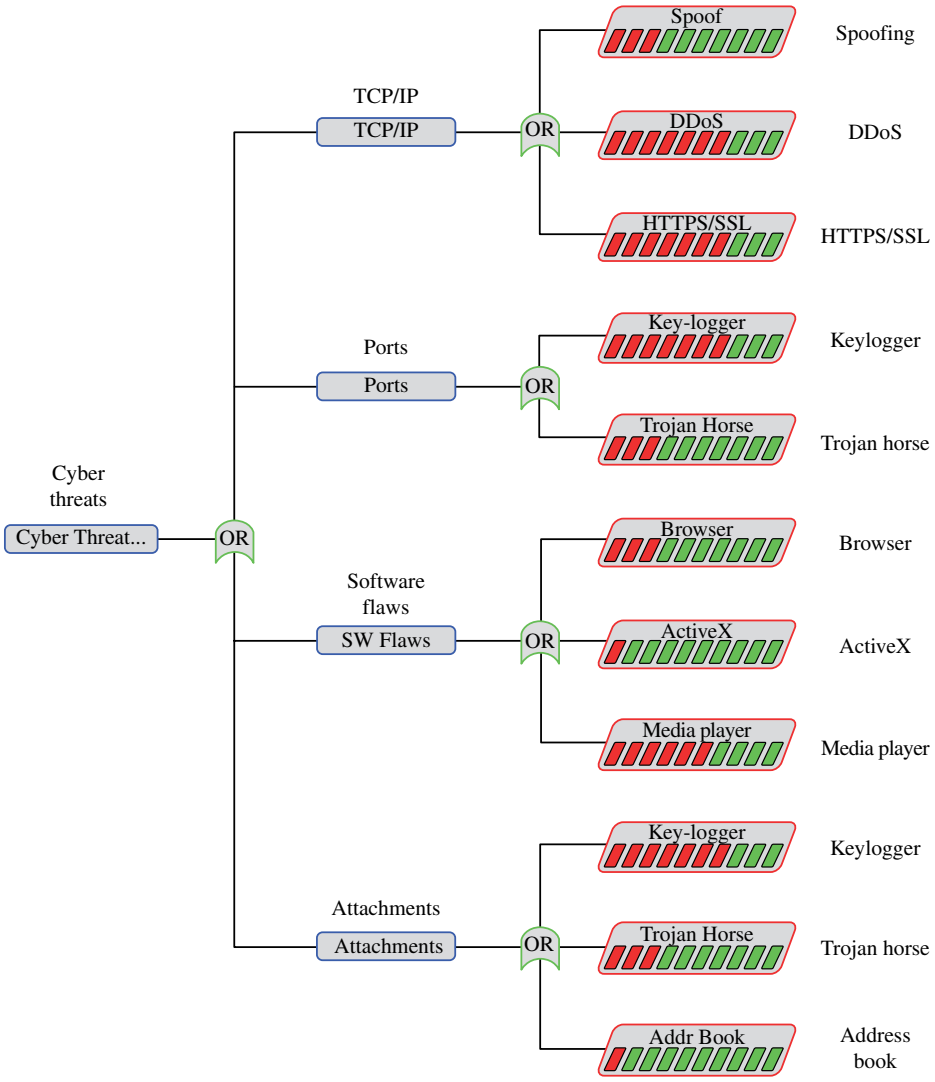


FIGURE 7.7 Common threat–asset pairs in a general fault tree of cyber threats.

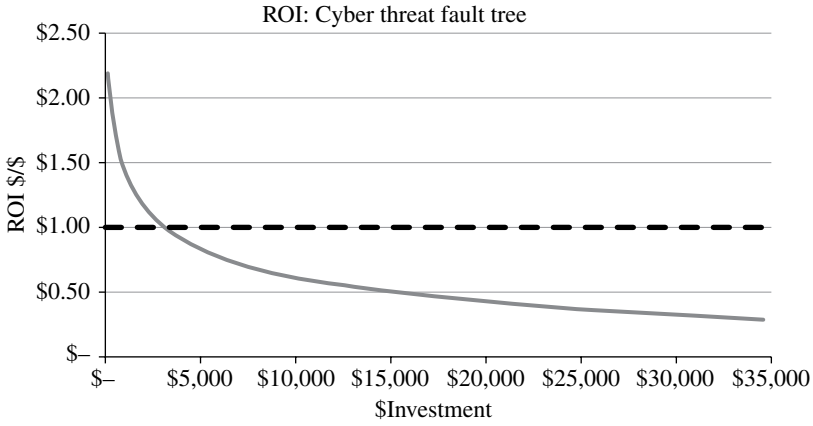


FIGURE 7.8 ROI analysis of the general fault tree model for the values shown in Table 7.1 suggesting the optimal investment is approximately $3000.

TABLE 7.1 Input values for the general fault tree model of Figure 7.7 are used to optimally allocate $3000 to reduce vulnerabilities

Name	Threat	Vulnerability	Elimination cost	Consequence
Spoof	100	10	2,500	10,000
DDoS	100	10	5,000	10,000
HTTPS/SSL	100	10	5,000	10,000
Keylogger	100	10	5,000	10,000
Trojan horse	100	10	2,000	10,000
Keylogger	100	10	5,000	10,000
Trojan horse	100	10	2,000	10,000
Keylogger	100	10	5,000	10,000
Trojan horse	100	10	2,000	10,000
Browser	100	10	2,000	10,000
ActiveX	100	10	1,000	10,000
Media player	100	10	4,000	10,000
Address book	100	10	1,000	10,000

to Chapter 6 to reexamine the impact an Internet infrastructure attack might have on individual IT installations. Protecting individual IT installations or PCs has almost no effect on protecting the Internet. Conversely, protecting a handful of major Internet hubs has an enormous impact on protecting individual IT installations and PCs.

The AS2000 Internet network of Figure 7.9 contains 2000 autonomous systems and 6107 links. It is highly structured with a power law distribution of links to nodes and a spectral radius of 45.8. The hub contains 388 connections, but mean connectivity is only 6.1 connections. Link robustness is very high and its experimentally determined number of blocking nodes is 247 (12.4%).

The AS2000 Internet has a small number of very highly connected servers and thousands of servers with only a handful of connections. The high spectral radius and large hubs make the AS2000 network a near-perfect superspreader of malicious software. This structure also explains why AS2000 can be fractured into disconnected pieces by removal of only 247 (12%) servers—most connectivity is vested in a handful of highly linked hubs.

The highly structured shape of the AS2000 network suggests an optimal protection strategy—harden the hubs and ignore small IT installations and individual computers. Invest heavily in the 247 blocking nodes because these autonomous servers hold the entire system together. In fact the following analysis shows that installing antiviral software on individual computers is ineffective and a waste of resources. On the other hand, hardening the servers in the top hubs of the Internet is very effective. For example, hardening a mere 40 of the AS2000 network's most connected servers is 3.7 times as effective as protecting individual desktops, PCs, and laptops.

Figure 7.10 shows the results of a number of cascade simulations carried out on the AS2000 network of Figure 7.9. Four simulations were performed with varying levels of infectiousness: (1) no protection, (2) randomly selecting 2% of the nodes for protection, (3) protecting 2% of the least connected nodes, and (4) protecting 2% (40) hub nodes and ignoring all the others. In all cases except for the last one, protection of 2% of the nodes was ineffective. But when the top 40 nodes (hubs) were hardened so they could not spread malicious software to adjacent neighbors through peering, the spreading was nearly halted altogether.

The fundamental resilience line for these four simulations on AS2000 shows that hub protection is 3.7 times more effective than any other strategy. Indeed, individual computer system (singleton) protection was no more effective than random protection. In other words, individual antivirus software installed on laptop and desktop computers is far less effective than hardening the top 40 AS servers in the Internet.

Effective Protection Strategy: *The highly percolated and structured Internet is protected against the spread of malicious software by hardening a very small percentage of its hubs, typically less than 2–3%.*

7.5.1 Blocking Node Analysis

If the 247 blocking nodes can be found, hardening them prevents the spread of malicious software from one component to an adjacent component. Computer viruses cannot jump from one island to another, if a blocking node prevents it. Blocking nodes do exactly that—they block further spreading of cascade faults. Perfect hardening of all 247 blocking nodes can stop spreading in its tracks. But how are these highly critical nodes found?

We can use the structure of a scale-free network to identify most of the blocking nodes. Scale-free networks concentrate connectivity in hubs. Hubs are the most likely to link components together. They are the strongest glue

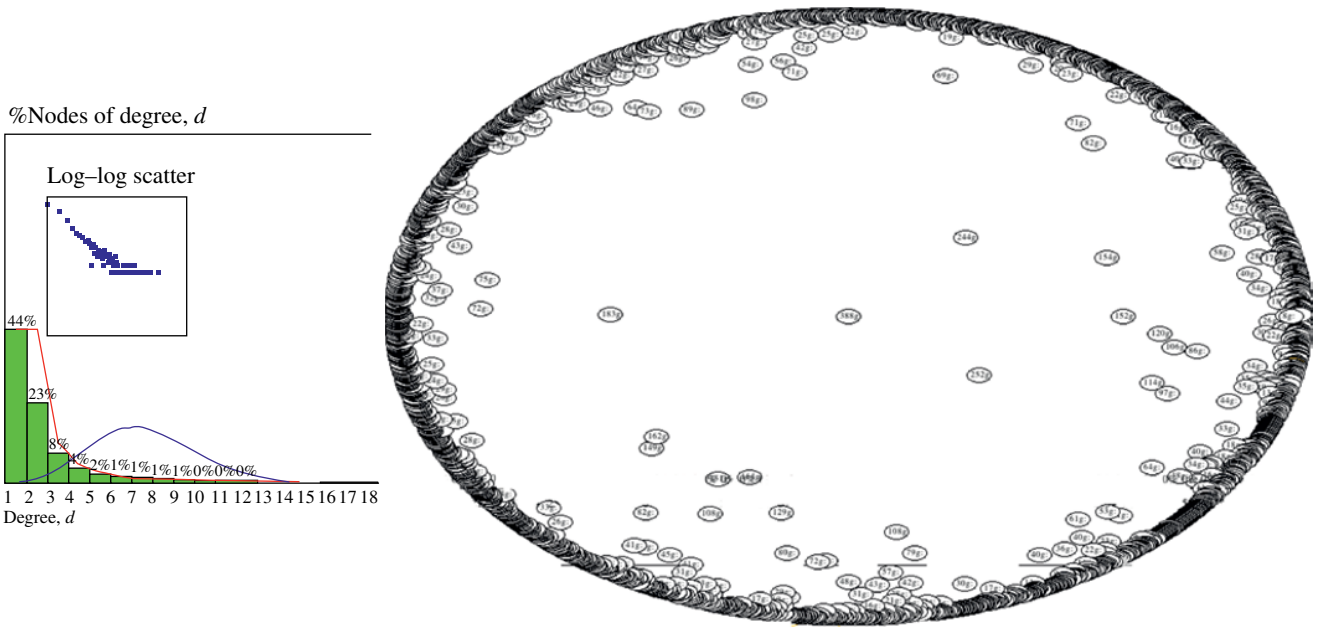


FIGURE 7.9 AS2000: The top 2000 autonomous systems of the global Internet form a near-perfect scale-free network with a 388-link hub at its center, a power law exponent of 1.93, and a spectral radius of 45.8. Links have been removed to more clearly show the nodes and their degree distribution (higher-degreed nodes are placed near the center).

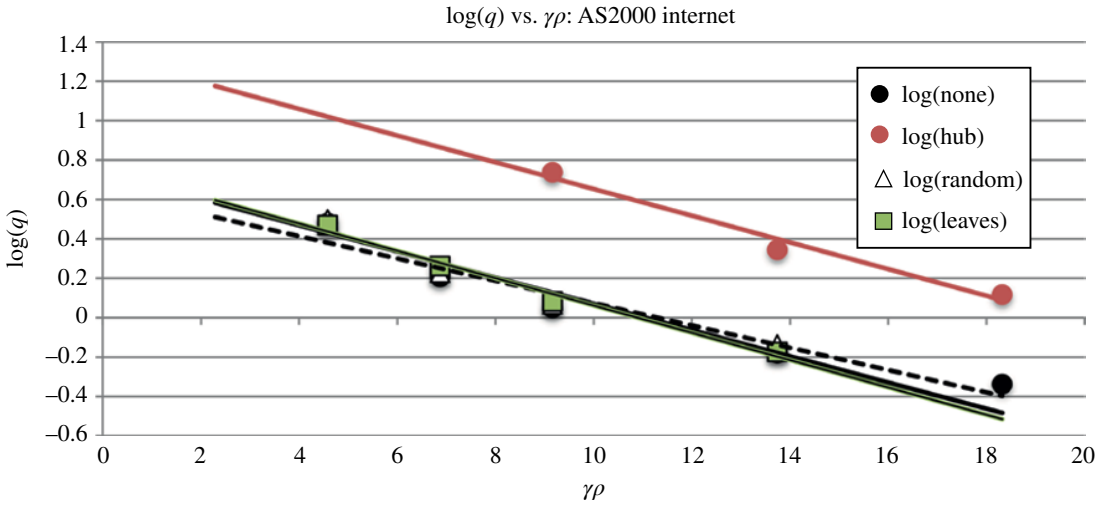


FIGURE 7.10 Result of four simulations shows hub hardening to be 3.7 times as effective as hardening individual computers.

holding the network together. Therefore, we need only rank the most connected 247 AS2000 nodes in descending order by connectivity to identify the most likely blocking nodes. In this case, 80% of the nodes have 2 or more links. Unfortunately, 88% must be eliminated. Rank ordering by connectivity is only a partial solution.

An exact method of finding blocking nodes requires a fast computer and brute-force algorithm. Each node is removed from the network one at a time. If it is impossible to trace a path from each of the removed node's neighbors back to itself, the node is a blocking node because its removal breaks the chain from neighbor back to neighbor. This algorithm must be repeated for every node in the network. Blocking nodes are the collection of nodes that break the chain of hops from each of their adjacent nodes back to itself.

Blocking Strategy: *Malicious software is stopped from spreading by hardening all the blocking nodes in the AS2000 network, because hardened blocking nodes separate the network into disjoint islands.*

7.5.2 Machine Learning Approach

Network analysis may be useful for halting the spread of malware, but it has little power to detect malware in the first place. In fact, malware detection is a difficult technical problem with many techniques and vendors claiming to be able to detect and block malware. The reality is much different. If detection was a solved problem, the Internet would not be under constant attack from criminals and state actors.

One promising approach is to use machine learning algorithms to detect malware by examining the signature of all data attempting to enter a computer system before it is allowed to enter. A machine learning algorithm is a computer program that has learned to find patterns in data. Suppose the data in this case is a stream of information attempting to enter a computer system. It might be a valid program or data, or it might be malware. How does the machine learning algorithm determine which is the case?

A signature is any piece of information that uniquely describes data and programs. Its purpose is to automate malware detection by matching signatures of "good" and "bad" streams of data against a database of signatures. Automated systems for detecting malware use either static or dynamic signatures. Static signatures are patterns such as program length, where the stream of data came from, and so on. If a stream contains at least one known pattern, it is classified as malware.

Dynamic pattern matching uses the dynamic structure of a program as a signature. Executing programs have somewhat unique structures as defined by their flowchart or call graph. A call graph, for example, is simply a hierarchical "organization chart" of a program's inner structure based on the pattern of jumps from one module to others, deep within the program's code. Thus, a call graph can be used as a dynamic signature.

A machine learning algorithm such as an artificial neural network can be trained to classify inbound programs and data as malicious or not, based on known signatures [5]. The artificial neural network reads every inbound stream of data and looks for patterns that match known malware patterns or signatures. That is, the artificial neural network classifies inbound data as malicious or not.

The future of malware countermeasures is automation, because humans are not fast enough or clever enough to detect and deflect malware as it spreads rapidly across the Internet. However, application of machine learning algorithms to the problem of malware detection is yet another step in the arms race between criminals and computer security professionals. It is unlikely to end soon.

7.5.3 Kill Chain Approach

Chapter 2 describes the kill chain framework for detecting and responding to malware attacks and suggests a fault tree model of risk due to malware penetrating the kill chain. This concept is also related to the notion of a trusted path in computer security—a topic developed in more detail in the next chapter. Recall the steps taken by a prototypical computer security attack defined by the kill chain framework:

1. Reconnaissance—The attacker looks for vulnerabilities.
2. Weaponization—The attacker develops one or more vehicles to deliver the payload, which may be a RAT, botnet, DDoS, or some new weaponized device or process.
3. Delivery—The attacker selects a delivery mechanism such as an email, browser flaw, thumb drive, and so on.
4. Exploitation—The attacker triggers the malware from afar through a command as in a botnet, a simple spear phishing telephone call to an employee, and so on.
5. Installation/spread—Once installed on a victim's machine, the malware may spread to other victims, for example, using the victim's address book.
6. Command and control (C2)—The attacker may issue commands from afar or automate the malware so it activates automatically on a certain date, such as New Year's Day.
7. Actions on objectives—Typically, the objective is data exfiltration, which involves collecting, encrypting, and extracting information from the victim environment.

Every exploit illustrates one or more steps of the kill chain framework. For example, consider a typical phishing exploit that almost any criminal can perform without much expertise—installation of a keylogger on a victim's desktop computer. Sally uses a computer at work to receive and communicate with co-workers by email. On Monday morning she receives an email from her IT support technician named Sam. The email asks Sally to update her system by downloading an attachment containing the latest version of the company-approved word processing software. Sally obeys and clicks on the attachment to start the downloading and updating process.

Unfortunately, the email is a fraud. By clicking on the attachment, Sally inadvertently installs a keylogger on her desktop computer. The keylogger begins to record her keystrokes and sends them to a botherder looking for passwords into the corporate computer system. Sally may not even realize she has assisted the virus and aided the hacker.

From a kill chain framework perspective, the exploit starts with reconnaissance. The hacker found Sally's name, email address, and job information from her Facebook.com page, which revealed that she works for a certain bank. Social network sites are a rich source of reconnaissance information for criminals.

Weaponization is even easier because keyloggers are easily purchased from the dark web for small sums of money. In this case, the hacker purchased a keylogger for $100 and

began sending it as an attachment to hundreds of victims with public Facebook.com accounts. The delivery method is a simple email with the keylogger malware as an attachment. Of course, the attachment is renamed so it appears to be an update of known software.

Sally triggers the keylogger by clicking on it and allowing it to download into her system. It immediately begins to record her keystrokes. The hacker is looking for passwords to the company database. When Sally logs into the bank's database, her password is sent in the clear to the hacker. Eventually, the criminal gains access to the bank's accounts and begins exfiltrating user information such as names, passwords, and home addresses.

The kill chain is completed when the hacker successfully decodes encrypted passwords downloaded from Sally's bank database. The hacker uses another dark web program for deciphering encrypted passwords that the criminal purchased for $250. The scheme is so successful, in that the hacker emails hundreds of people whose email addresses and names are found on Facebook.com, Twitter.com, and Instagram.com.

7.6 ANALYSIS

How likely is it that cyber exploits will be used by terrorists in the future? In a report released immediately after 9/11, Michael Vatis, Director of the Institute for Security Technology Studies at Dartmouth University, claimed cyber attacks are highly correlated with physical attacks and terrorism:

> In the Israel/Palestinian conflict, following events such as car bombings and mortar shellings, there were increases in the number of cyber attacks. Subsequent to the April 1, 2001 mid-air collision between an American surveillance plane and a Chinese fighter aircraft, Chinese hacker groups immediately organized a massive and sustained week-long campaign of cyber attacks against American targets. [6]

Vatis argues that cyber attacks immediately accompany physical attacks and they increase in volume, sophistication, and coordination. He also correlates these attacks with high-value targets.

Thus far, nobody has died as a direct result of a cyber attack. In fact, James Lewis argues that the threat of cyber attacks from terrorists has been exaggerated:

> Digital Pearl Harbors are unlikely. Infrastructure systems, because they have to deal with failure on a routine basis, are also more flexible and responsive in restoring service than early analysts realized. Cyber attacks, unless accompanied by a simultaneous physical attack that achieves physical damage, are short lived and ineffective. However, if the risks of cyber-terrorism and cyber-war are overstated, the risk of espionage and cyber crime may be not be fully appreciated by many observers. This is not a static situation, and the vulnerability of critical infrastructure to cyber attack could change if three things occur. Vulnerability could increase as societies move to a ubiquitous computing environment when more daily activities have become automated and rely on remote computer networks. The second is that vulnerability could increase as more industrial and infrastructure applications, especially those used for SCADA (Supervisory Control and Data Acquisition), move from relying on dedicated, proprietary networks to using the Internet and Internet protocols for their operations. This move to greater reliance on networks seems guaranteed given the cost advantage of Internet communications protocols (Transmission Control Protocol/Internet Protocol), but it also creates new avenues of access. These changes will lead to increased vulnerabilities if countries do not balance the move to become more networked and more dependent on Internet protocols with efforts to improve network security, make law enforcement more effective, and ensure that critical infrastructures are robust and resilient. [7]

Nonetheless, cyber threats still exist and are responsible for major financial losses. They will continue to be a threat for as long as computer systems have flaws. And flaws are expected to remain a part of this sector for a long time, because fallible humans build information systems. Moreover, cyber attacks are highly asymmetric, meaning they are cheap and easy to apply. As society continues its adoption of all things Internet, the consequences of damage to CIKR will also continue to increase.

Moreover, the era of script kiddies and criminal blackhats is phasing out as the era of state-sponsored hacking and cracking is phasing in. The future of cybersecurity may be automation, but the future of hacking and cracking is human ingenuity. State-sponsored hacking is more likely than ever to be run by armies of humans pitted against machines. And these armies are weaponizing malware that will be integrated into kinetic operations in the next conflict between nations.

7.7 EXERCISES

1. What is the precise definition of a virus?
   a. A malicious self-activating program
   b. An email attachment
   c. A malicious user-activated program
   d. A malicious program that travels via the Internet
   e. A flaw in TCP/IP
2. What is the precise definition of a worm?
   a. A malicious self-activating program
   b. An email attachment

c. A malicious user-activated program
d. A malicious program that travels via the Internet
e. A flaw in TCP/IP

3. What is a Trojan horse?
a. A malicious self-replicating program
b. An email attachment
c. A malicious program disguised as a safe program
d. A malicious program that travels via the Internet
e. A flaw in TCP/IP

4. Software patches are:
a. Often not installed on vulnerable systems
b. A defect that opens a computer to attack
c. A software developer's repair kit
d. A kind of virus
e. A way to repair an open port

5. Which one of the following is an old method of breaking into a computer?
a. War dialing
b. War driving
c. SIS
d. SOS
e. SOB

6. A zombie is a:
a. Defective computer
b. Network of dead computers
c. Innocent participant in a DDoS attack
d. Computer cyber thief
e. Black-hat cyber criminal

7. The AS network is characterized by:
a. Giant secure connected component
b. Grand secure connected component
c. Giant strongly connected component
d. 2000 nodes
e. Very high level of percolation

8. Typical TCP/IP exploits are:
a. Focused on attachments
b. Focused on design flaws
c. Phishing expeditions
d. Buffer overflow attacks
e. Fixed by installing the latest software patches

9. Which of the following is NOT a critical node in the Internet?
a. Amazon.com Web services data center
b. Typical desktop or laptop computer
c. The MAE-West hub in San Jose, CA
d. The hub at the Chicago NAP
e. The root servers at a.root-servers.net

10. As far as we know, the first cyber worm was launched in:
a. 1998
b. 2001
c. 9/11/01
d. 1988
e. 1984

11. A malicious program can be a macro that travels by:
a. A Word or Excel document
b. Downloading itself through a buffer overflow attack
c. Downloading itself through spyware
d. Attaching itself to an email
e. Embedding itself in a Trojan horse

12. A DDoS attack:
a. Floods a victim computer with a huge number of messages
b. Uses email to send fake messages to users listed in address books
c. Is a special kind of worm
d. Uses macros to travel
e. Blocks ports

13. The Bank of America ATM network was temporarily stalled by:
a. A SYN flooding DDoS attack
b. SQL Slammer
c. MS-Blaster
d. Klez
e. The Morris worm

14. The Whitehouse of the United States was attacked in 2002 by:
a. Bugbear
b. MafiaBoy
c. Code Red
d. Microsoft IIS
e. Changing the DNS server address

15. What are ports?
a. Input/output channels through which network information flows
b. Vulnerable flaws in the global Internet
c. 65,535 doors
d. TELNET input
e. FTP input/output

16. In a buffer overflow attack:
a. A program enters a computer as if it were data
b. A malicious program travels through ports
c. A worm exploits FTP
d. A worm exploits port 21
e. An operating system is exploited

17. Most DDoS exploits use:
a. SYN flooding
b. Zombies
c. Web servers
d. Microsoft Windows flaws
e. Routing tables

18. MafiaBoy caused $1.7 billion in financial loss. How much was the fine?
a. $10,000,000
b. $1,000,000
c. $10,000

d. $250
e. None

19. Who turned in the WikiLeaks traitor?
    a. Pfc. Bradley Manning
    b. Dark Dante
    c. Kevin Poulsen
    d. *The New York Times*
    e. Adrian Lamo
20. Which of the following is the best protection strategy for reducing the spread of malicious software throughout the information technology sector?
    a. Harden the Internet's hubs
    b. Install antivirus software on all desktops and laptops
    c. Install antivirus software on all end-user devices including smartphones
    d. Enacting stronger laws
    e. Redesigning the TCP/IP protocol

7.8 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. If computer security risk is due to flaws in software, why do software developers allow them to exist and why don't they get repaired before a zero-day exploit is discovered?

B. TCP/IP is well known to be non-secure. Why isn't it fixed or replaced by a secure protocol?

C. The topology of the AS-level Internet is known, which means we also know where the super-spreaders are located. Propose three policies or technologies that can be used to reduce or stop the spread of malware via super-spreaders. Why are super-spreaders the proper targets of your policies and technologies?

D. The family tree shown in Figure 7.3 suggests something about the evolution of malware. What does it suggest for the future of malware?

E. The Internet is a monoculture for a number of reasons. What are some reasons, and suggest how to reduce Internet vulnerability by diversifying its protocols. Which protocols do you think are the easiest to diversify?

REFERENCES

[1] Dunlevy, C.J. Protection of Critical Infrastructures: A New Perspective. *CERT Analysis Center*, 2004. Available at http://www.cert.org. Accessed June 20, 2014.

[2] Harris, S. The Cyberwar Plan, Not Just a Defensive Game, *National Journal*, November 13, 2009. Available at http://www.nextgov.com/nextgov/ng_20091113_1728.php. Accessed June 20, 2014.

[3] Shane, S., Nicole, P., and Sanger, D.E. Security Breach and Spilled Secrets Have Shaken the N.S.A. to Its Core. *New York Times*, November 12, 2017. Available at https://www.nytimes.com/2017/11/12/us/nsa-shadow-brokers.html. Accessed June 20, 2014.

[4] Abu-Ghazeleh, N., Ponomarev, D., and Evtyushkin, D. How the Spectre and Meltdown Hacks Really Work, *IEEE Spectrum*, 56, 28, February 2019, pp. 42–49. Available at https://spectrum.ieee.org/computing/hardware/how-the-spectre-and-meltdown-hacks-really-worked. Accessed June 20, 2014.

[5] Hill, F. *Graph-Based Malware Classification using Machine Learning*. Ruhr-Universität Bochum, Master's Thesis, September 2017.

[6] Vatis, M. *Cyber Attacks During the War on Terrorism: A Predictive Analysis*, Hanover, NH: Institute for Security Technology Studies, September 22, 2001. Available at http://www.ists.dartmouth.edu/ISTS. Accessed June 20, 2014.

[7] Lewis, J. A. *Assessing the Risks of Cyber Terrorism, Cyber War and Other Cyber Threats*, Washington, DC: Center for Strategic and International Studies, December 2002. Available at http://www.csis.org. Accessed June 20, 2014.

8

INFORMATION TECHNOLOGY (IT)

Information technology (IT) sector vulnerability is defined by the US Congress as "the vulnerability of any computing system, software program, or critical infrastructure, or their ability to resist, intentional interference, compromise, or incapacitation through the misuse of, or by unauthorized means of, the Internet, public or private telecommunications systems or other similar conduct that violates Federal, State, or international law, that harms interstate commerce of the United States, or that threatens public health or safety."[1] For our purposes, cybersecurity is the study and practice of securing assets in cyberspace—the world of computers and computer networks. Cybersecurity is more than defending against viruses and worms, as described in Chapter 7. It encompasses *information assurance* in *enterprise computing*.

This chapter surveys the policies and technologies of securing information and the IT systems that process information—the IT sector. The phrases *cybersecurity* and *IT sector security* will be used interchangeably. The essence of IT security centers on the notion of *trusted computing*—a *trusted computing base* (TCB) containing hardware and software, plus *trusted paths* (TP) between and among various computing bases. In layman terms, this means encapsulating hardware, software, and data in a protected zone and protecting communication transactions between and among users.

The rules of trusted computing have been known for many decades, so what is the problem? For the most part, trusted computing depends on human processes as much, if not more, than on technology. IT security is a human process problem. Moreover, securing an enterprise computing system is not easy or inexpensive. Even if IT owners and operators are aware of a threat, there is an economic disincentive to respond to the threat. An unfortunate example of this occurred in 2017 when hospitals in England neglected to upgrade outdated operating systems even when it was widely known they were vulnerable to ransomware attacks. Had they updated to the latest version of Microsoft's operating system, they would not have been affected. Failure to patch system software is a common mistake, but it is not without a reason. It is time consuming and expensive.

IT security introduces inconveniencies and requires additional effort. It requires eternal vigilance. It is an added expense. It introduces an inconvenience to users. Therefore, IT security policies must strike a balance between ease of use and costly protection of users and user's data.

Specifically, this chapter discusses the following:

- *Principles*: There are five fundamental principles of IT sector security. The IEEE X.509 standard specifies four: authentication, information integrity, information confidentiality, and non-repudiation of ownership. The fifth is availability—or freedom from DDoS attacks. Authentication is typically achieved by passwords, and integrity, confidentiality, and non-repudiation are achieved by encryption. Installing proxy servers and automatic detection and deflection software deflects DDoS attacks.

[1]HR 4246 introduced into the 106th Congress 2nd session, April 2000.

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

- *Policies*: Cybersecurity involves a wide range of information assurance policies and practices including but not limited to the prevention of loss of access to information, loss of data, and loss of security associated with IT and human information-handling processes. For example, a typical policy of any large IT operation is to update passwords twice per year and to use 16-character, or larger, passwords.
- *Trusted computing*: Secure IT systems are based on a TCB and enforce the use of TP through a network of IT components and human users to ensure the security of information stored and processed by the IT system.[2] To be secure, all IT processes must run within a TCB and communicate via a TP.
- *Components*: The major components of a TCB and TP are firewalls, proxies, *intrusion detection systems* (IDS), encryption, public key encryption (PKI), and policies that enforce a certain level of user hygiene. Security is not an absolute "secure" or "not secure" decision, but rather a trade-off with other factors.
- *Encryption*: There are two basic types of encryption: *symmetric* and *asymmetric*. Symmetric encryption is used to secure information between trusted parties; asymmetric is used to secure information between anonymous parties. Both kinds of encryption have political implications because ciphers have historically been viewed as a form of munitions. Law enforcement wants backdoor access to encrypted data, and privacy advocates want strong encryption that cannot be cracked by anyone.
- *DES and AES*: Symmetric ciphers such as the Standard DES, Triple DES (3DES), and AES evolved out of the Lucifer project started in the 1960s by IBM, but the ideas go further back—perhaps as far as the classified work of the British during World War II. As computers get faster, symmetric codes are broken, requiring longer and longer keys. One major disadvantage of symmetric codes is that they are symmetric, which leads to their vulnerability and ultimately cracking.
- *Current standard encryption*: AES is the latest symmetric code to be standardized by the US National Institute of Standards and Technology (NIST) (2002), and besides being strong (256-bit keys), it is suitable for small computers such as those used in SCADA systems. 3DES and AES have been adopted by the US federal government and are required in order for an IT system to be FIPS compliant.[3] However, symmetric encryption assumes all parties know the key, which sets up the key distribution problem.
- *Diffie–Hellman cipher*: Asymmetric ciphers rediscovered by Diffie and Hellman in 1976 solve the key distribution problem by using two keys: a public key to encode and a private key to decode. The private key is not shared; hence it is less vulnerable to cracking. The RSA algorithm implements the ideas of Diffie and Hellman and makes it possible to authenticate users (digital signatures) as well as protect the privacy of both sender and receiver. Rivest, Shamir, and Adleman invented a practical method of performing the Diffie–Hellman algorithm, known as the RSA algorithm, which is the foundation of Internet security.
- *Certificates*: PKI authenticates the identity of users by assuring that the sender is who he or she claims; guarantees the integrity and security of the message by assuring that it has not been modified by an intermediary; assures privacy by making sure the message is decodable only by the intended recipient; guarantees authentication, security, and privacy is enforceable by assuring that the message is signed by the verified parties; and guarantees non-repudiation by assuring that both parties cannot disavow or deny involvement with the transaction.
- *FIDO*: Fast IDentity Online (FIDO) is an industry standard public key protocol for securely logging in to a trusted site. It uses a fast and low-friction "certificate-like" algorithm to automate encrypted communication between a user and a Web site. Standardized in 2019 by the W3C as WebAuth, the protocol eliminates the need for two-factor authentication (2FA) and passwords in general.
- *Passwords*: Passwords are likely to be replaced by WebAuth protocols that do not require passwords, but when passwords are still needed, care must be taken to avoid password hackers. Hash functions scramble passwords so they are not hackable even if a criminal exfiltrates a server's password file. Good hygiene requires that users select non-obvious passwords.
- *Strategy*: Cybersecurity will improve when the following information infrastructure improves: TCP/IP encryption of source/destination addresses; vendors remove software flaws; software defaults are configured for the highest level of security; users are better informed and trained to prevent security breaches; organizations adopt stronger standard operating procedures; consumers demand better IT security; and vulnerability and risk analysis are standardized and used routinely.
- *Incomplete knowledge*: More research needs to be done to make software virus-proof, reduce software errors that hackers can exploit, standardize risk analysis including the use of quantitative techniques, and develop new methods to analyze cascade effects, predictive methods, and recovery.

[2]A trusted path is a mechanism by which a person using a terminal can communicate directly with the trusted computing base (TCB). The trusted path can only be activated by the person or the TCB and cannot be imitated by untrusted software.

[3]FIPS is the Federal Information Processing Standard.

- *Quantum cryptography*: Quantum computing is powerful enough to crack the RSA encryption standard that is the basis of all PKI. Soon, the RSA algorithm will no longer be sufficient. However, quantum key distribution solves the problem created by quantum codebreakers. The future of IT security is in quantum computing and its cousin, quantum communication.

8.1 PRINCIPLES OF IT SECURITY

The IT sector is notoriously non-secure, and yet the principles for a secure IT infrastructure have been known for decades. The IEEE X.509 and RFC 2459 (1999) standards define cybersecurity in terms of four fundamental principles:

- Authentication: Ability to verify authenticity of users and data.
- Integrity: Ability to guarantee document or message has not been altered.
- Confidentiality: Ability to conceal the content of documents and messages.
- Non-repudiation: Inability to deny authenticity: non-concealment of ownership.

Authentication means the identity of a user is known—typically through the use of a username and password, but also through various biometric identification technologies such as fingerprints or voice recognition. *Integrity* means that email, attachments, and documents such as spreadsheets, photos, audio, and text arrive at their destination unaltered. *Confidentiality* means it is possible to store and transmit information without prying eyes "in the middle." Confidentiality is typically achieved by encryption. *Non-repudiation* means the sender cannot deny sending the document or message. For example, a message cannot be spoofed. Email from whitehouse.gov actually came from the Whitehouse, and contracts from your attorney actually came from your attorney.

A fifth goal is designed to fend off DDoS attacks. Availability is the ability of an IT system to be available for use. A DDoS attack prevents access and use of an IT system much like a traffic jam on a busy avenue prevents the use of the roadway. While availability is essential, it is relatively easy to circumvent. IT owners and operators know how to detect and divert DDoS attacks through various means.

The IEEE X.509 goals may seem simple on the surface, but they have proven to be difficult to implement in practice. One of the major barriers has been the very people they were meant to protect—consumers. Authentication requires the user to remember relatively long passwords; integrity, confidentiality, and non-repudiation require strong encryption and secure key escrow accounts. Users have been loath to adopt these technologies because they are inconvenient and unfriendly. Nonetheless, progress has been made over the past two decades as consumers opt for security even at the expense of convenience. The major technology responsible for this progress is called *public key infrastructure* (PKI), and the basic technology that PKI depends on is the *certificate authority* (CA).

In general, PKI is a combination of public key encryption and a hierarchical storage system based on certificates. A certificate is a digital document containing a user's authentication and encryption information. At a minimum, a digital certificate contains:

- The user's name
- The user's public key
- The public key's expiration date
- The CA that issued the certificate
- Digital signature of the CA

Digital certificates are like many common identification cards such as a birth certificate, driver's license, or credit card. But digital certificates are entirely digital and live in a CA database. CAs are hierarchical—local certificates are verified by higher-level authorities.

The general idea of cybersecurity is to enclose an IT system in a protected shell called the *demilitarized zone* (DMZ). All transactions occur within the DMZ along TP—nodes and links that are guaranteed to be secure as defined by the four security principles: authentication, integrity, confidentiality, and non-repudiation. Trusted systems manage trusted paths, which in turn deliver trusted IT services to users. Trusted systems, however, are difficult to implement.

8.2 ENTERPRISE SYSTEMS

An enterprise system is an IT system that is used by an enterprise—corporation, government agency, school, military command, and so on—regardless of size. Because an entire organization depends on it, an enterprise system demands high availability, data integrity, reliability, and security. Enterprise systems differ from other systems mainly due to their size, but they also differ in terms of their applications and performance. For example, typical enterprise systems run payroll, inventory, and Web and streaming services.

A desktop computer may be a member of an enterprise system, but it is not an enterprise system on its own. Enterprise systems span entire organizations and provide a stable core of hardware and software components that support the mission of an organization. They consist of computers of all sizes, networks for connecting them, and software for making them useful. Enterprise systems must scale to large numbers of users.

The centralized enterprise systems of the 1960s and 1970s returned in the 2000s in the form of cloud computers. A large central mainframe typically supported hundreds of remote users through a relatively low bandwidth communication

link. But by the 2000s, the rise of the Internet, low-cost wireless smartphones, and handheld apps reignited the demand for centralized enterprise systems. Instead of a single mainframe, the centralized system consisted of thousands and even millions of computers working in unison. The collection of cooperating computers located in remote parts of the world became known of as the cloud. Cloud computing is a highly mature collection of communication, massively parallel and cooperating computer technology, accessed by millions of users at once.

Aggregation of computation and data in the cloud has exacerbated the problem of computer security because the cloud is a single point of failure and "puts everyone's eggs in one basket." The increase in security requirements have not been offset by increased computer security technology nor the ability of organizations to protect consumer's privacy. Cloud computing has made computer security more important, but not stronger.

Unfortunately, it is theoretically impossible to determine whether or not an enterprise system is secure.[4] The best we can do is institute policies that diminish the likelihood that an enterprise system is compromised—either maliciously or inadvertently. Given the current state of computer security, cybersecurity is largely a practice rather than an exact science.

Generally, the goal of cybersecurity is to protect an enterprise system from loss of service, loss of data, and loss of security. *Loss of service* typically means the system is down, slow, or otherwise unable to respond to its users. *Loss of data* means that information is lost, and *loss of security* means the system has been compromised, either by a break-in or lack of proper controls such as access rights (failed password, user privileges, etc.).

8.2.1 Loss of Service

Loss of service can occur in at least three ways: power failure, telecommunications failure, and a denial-of-service attack. Power and telecommunications failures may be accidental or perpetrated incidents. A denial-of-service attack is an exploit perpetrated by an attacker.

Of course, there are many ways for loss of service to occur, such as malfunction of equipment and software defects that cause the enterprise system to stop. For example, power outages can be mitigated by backup power, and telecommunications outages can be mitigated by redundancy. Software defects can be partially mitigated by updates and patches, and databases can be backed up. DDoS exploits are detected by IDS placed between the enterprise system and the outside network. Each analysis must be tailored to the enterprise system under investigation.

[4]Deciding whether or not a computer system is secure has been shown to be impossible, by mathematical logic. Consider the following paradox: Tom says, "Sally always tells the truth," and Sally says, "Tom always lies." Is Sally lying, now? It is impossible to decide. In a similar fashion, system security can be shown to be undecidable.

8.2.2 Loss of Data

Loss of data can occur for a number of reasons: a file might be inadvertently deleted, a virus might be responsible for file deletions, or the deletion might be the result of an exploit that uses a flaw in an application such as Microsoft Excel, Oracle database, or human relations management software. Backing up databases and storing the backup off-site typically mitigate data loss.

For example, a break-in made possible by a clear password file may result in a malicious act such as an important file being deleted, but if there is a backup, then the file can be restored. Thus, a backup policy can assure the security of information even when files are deleted. How often should the enterprise make backup copies?

An application may be vulnerable to an attack or inadvertent loss of data because of a flaw in the application software. For example, a malicious attachment may be downloaded through a browser that operates in the clear or by allowing a malicious certificate to access a worker's desktop computer. If the malware successfully deletes the user's files, it can also spread throughout the enterprise and delete other files as well.

8.2.3 Loss of Security

Loss of security is what most people think of when they think of cybersecurity. This category includes a vast number of faults, described in greater detail in Chapter 7. Password violations are the most prevalent type of exploit. They occur because users fail to protect them, the enterprise system itself fails to protect the password file by encrypting it, or the enterprise system implements a weak PKI system.

Even 2FA is vulnerable to an SS7 attack as described in previous chapters. Many passwords are lost due to phishing—a social engineering exploit that coaxes a user into giving his or her password to a hacker. As enterprise systems require longer and longer passwords, lack of convenience sets in, reducing the effectiveness of extremely long passwords.

War dialing is an old-fashioned way to discover passwords. War dialing is automated cracking software that systematically dials phone numbers until a computer answers and then systematically tries all the words in the dictionary until one of them works. This is why passwords should be nonsense strings of characters including numbers and special symbols.

Login and password strings can also be obtained by recording keystrokes of the user (keylogging) or observing traffic over the network connecting the user to the Internet. If the user's login and password are transmitted in the clear, instead of encrypted, a *man-in-the-middle attack* may be used to get the user's password and username. To prevent this, a browser should always connect with the enterprise system through an HTTPS server using SSL (Secure Sockets Layer) or TLS (Transport Layer Security) encryption.

Loss of security can also occur because of a worm attack that succeeds in entering a victim's enterprise computer and then spreading to users connected to the enterprise computer. As described in Chapter 7, one such exploit begins with a buffer overflow, which succeeds because the victim has not installed the latest patch and the attacker has found an open port. A worm that achieves enough privileges to bypass the operating system's defense can take control of the entire enterprise system.

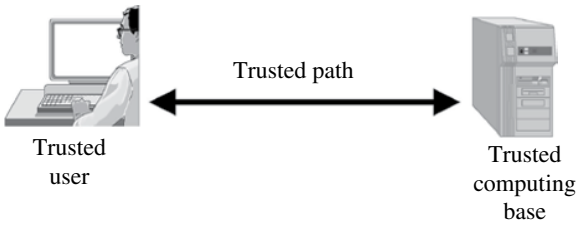


FIGURE 8.1 The architecture of a trusted computing base (TCB) consists of secure users, computers, and paths connecting them.

8.3 CYBER DEFENSE

Cyber defense is more a matter of policy than technology. On the one hand, security is costly and inconvenient for users. On the other hand, defense is necessary—at some level—in order to secure the information managed by an enterprise system. Generally, cybersecurity policy will end up reflecting many compromises between assurance and convenience. This leads to a question, "what is the minimum security, possible?" This section surveys a set of minimum policies for ensuring a *basic* or *foundational* level of cybersecurity called a *trusted computing base*.

Definitions of TCB vary, but for our purposes a TCB is the totality of protection mechanisms within a computer system—including hardware and software—that is responsible for enforcing a security policy. It creates the most fundamental protection environment possible along with some additional user services required for a trusted computer system. The ability of a TCB to correctly enforce a security policy depends solely on the mechanisms within the TCB and on the correct input of parameters by system administrative personnel (e.g. a user's clearance) related to the security policy.

Figure 8.1 shows a TCB made up of a TCB system and a TP between the computing system and user. The user will typically be a person sitting at a desktop, laptop, tablet, or smartphone connected to an enterprise system through the Internet. The TP will typically be a secured Internet connection.

The core component of a TCB is an enterprise server running behind a firewall that establishes a security zone called the DMZ (see Fig. 8.2). The DMZ forms a protective shell surrounding the components necessary for enforcing security policies such as user authentication, encrypted data, and access privileges. Security is guaranteed within the DMZ.

Outside of the DMZ is a TP connecting users to the DMZ. While a TP can be implemented by any kind of network, our example uses an encrypted Internet connection. In addition, a TCB must ensure that the users are who they say they are. The identity of the system's end users must be authenticated, usually by employing a user login and password. Of course, mobile devices such as tablets and smartphones often employ biometric methods of authentication such as fingerprints and voice pattern recognition.

The goal of this architecture is to establish a TCB made up of the components protected within the DMZ, a TP between user and DMZ, and a trusted user. Note that the TP follows closely the concept of a kill chain, because the kill chain analysis likely centers on TP. That is, an exploit by human or malware is likely to attack the TP, because it leads to sensitive information and control. Recall the seven rungs on the kill chain:

1. Reconnaissance
2. Weaponization
3. Delivery
4. Exploitation
5. Installation/spread
6. Command and control (C2)
7. Actions on objectives

The list of mechanisms at the bottom of Figure 8.2 suggests ways to ensure security of the TCB, TP, and authenticated users. Short explanations of common mechanisms for enforcing the TCB and TP follow.

8.3.1 Authenticate Users

Starting from the user's perspective, the first step in establishing a TP is to verify the authenticity of the users. The login/password mechanism currently used by most enterprise systems is perhaps the simplest. But more sophisticated biometric techniques should be used if a higher level of trusted computing is desired. For example, 2FA using a callback mechanism such as relaying permissions via an offline email account or telephone number may be employed to increase the strength of authentication. Selection of authentication technology is an example of a policy decision that will affect the implementation of cybersecurity.

Suppose, for example, a certain enterprise is required to guarantee the security of classified information. In this case, it is likely that a login/password mechanism will be inadequate for user authentication. Instead, users might be authenticated using retinal scanning, temporal passwords with a 30 s lifetime, and smart cards containing very long keys. Access to classified data is likely to require cryptographic keys such as in PKI described in Section 8.1.

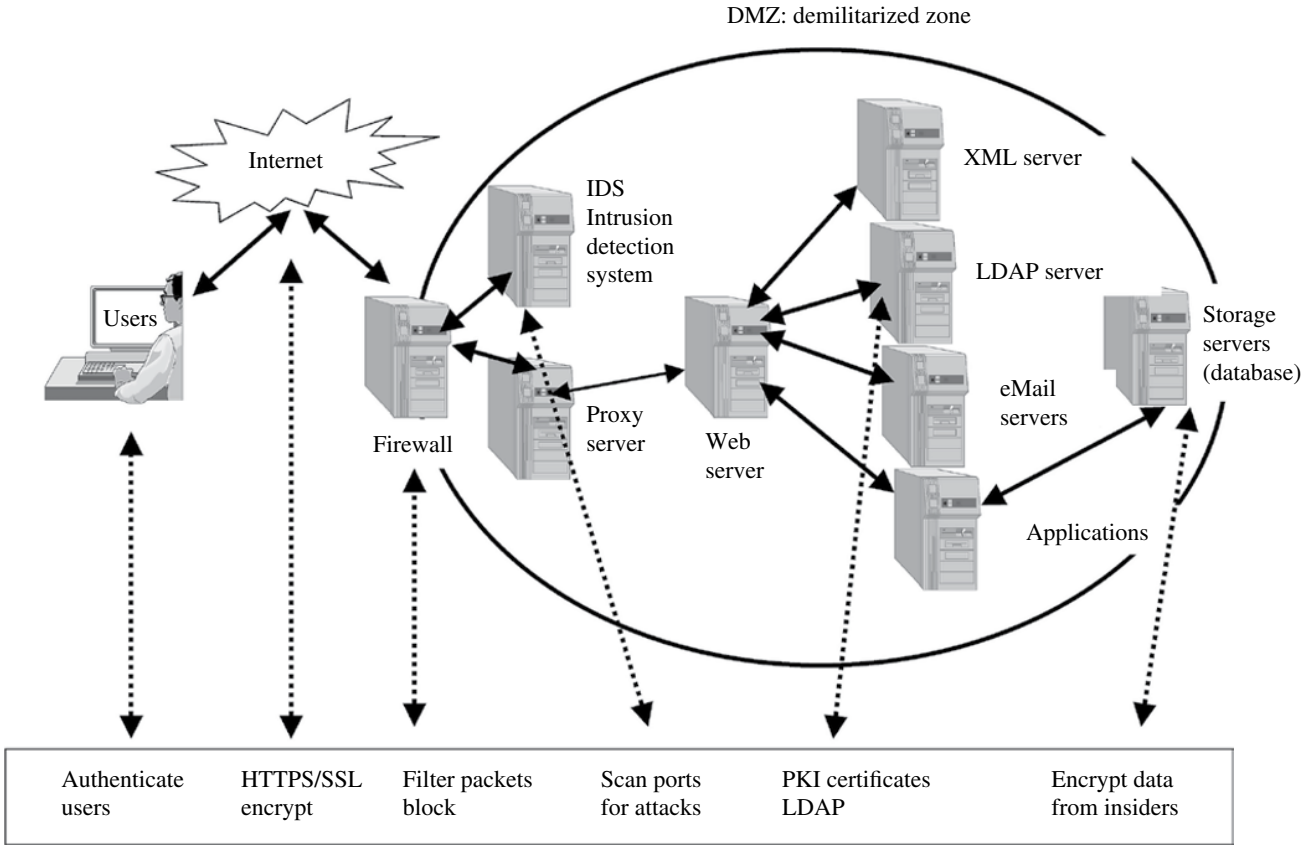


FIGURE 8.2 A detailed view of a typical TCB and the security technologies in common use by enterprise computing systems.

8.3.2 Trusted Path

The TCB needs a TP that connects users to the information they need to do their jobs. The purpose of a TP is to guard against man-in-the-middle attacks or fraudulent impersonation of valid users. Using a naval metaphor, we not only want to protect the ports but also ship on the high seas. Similarly, cybersecurity aims to protect information stored in the core of the TCB as well as information that is in transit between users and the TCB. This means protecting data at rest and data in transit.

The most elementary means of protecting communication links is to use a browser that supports SSL encryption, also known as TSL. SSL/TLS requires HTTPS (Hypertext Transport Protocol Secure) running on the enterprise server. SSL/TLS encrypts each session so that an intercepted communication cannot be hacked. E-commerce sites should always use the SSL/TLS/HTTPS combination to provide a TP for their online customers—especially when account numbers and personal data are transmitted. Credit card information should be encrypted, for example, and delivered by HTTPS to the e-commerce store. The SSL/TLS protocol is an integral part of the PKI standard established by IEEE X.509 in 1999.

SSL/TLS implements a modest level of cybersecurity. A stronger form is called virtual private network (VPN).[5] Recall that each TCP/IP packet is transmitted in the clear, meaning that a man-in-the-middle attacker can see both source and destination addresses. While the data may be encrypted, the remainder of the packet is not. IP version 6 (Ipv6) supports encryption of the TCP/IP packets, themselves, but less than half of all Web sites deploy Ipv6.

A VPN conceals not only the contents but the sender and receiver's identity as well. The "V" in VPN stands for "virtual," which means that virtual source and destination addresses are used in place of real addresses. To get through a firewall, these virtual addresses must be recognized and translated back into their real address equivalents. This is called IP *tunneling*, or VPN tunneling, because the VPN establishes a "tunnel" through the firewall. Tunneling involves establishing and maintaining a logical network connection with possibly intermediate hops. A VPN allows corporations to establish a proprietary network on an open public network such as the Internet.

[5]Virtual private network. A network that uses the Internet rather than leased lines for connections. Security is guaranteed by means of a *tunnel* connection in which the entire TCP/IP packet (content and addresses) is encrypted and encapsulated.

The bottom line is this: a VPN can be constructed on top of Ipv4 or Ipv6, making the TP much more secure—even though data travels over the open Internet. This can be costly in terms of hardware and software, and it can slow down a network because of the translation between virtual and real addresses. So the trade-off is cost, speed, and convenience versus enhanced security. Thus the decision to use a VPN must be a policy decision.

8.3.3 Inside the DMZ

Once inside the DMZ of Figure 8.2, implementation of cybersecurity becomes a more complex and sophisticated challenge, because a successful hack into the DMZ can have disastrous repercussions. If the DMZ is compromised all users and data are compromised. The question is, "what are the minimum mechanisms for achieving a minimally secure DMZ?"

A typical minimum set of mechanisms for assuring the security of a DMZ are the following:

- Firewalls.
- Proxies.
- IDS.
- Secure Web servers.
- Secure XML servers.
- Lightweight Directory Access Protocol (LDAP) servers.
- PKI software and policies for enforcing the TCB.

The first line of cyber defense is the firewall. A *firewall* is a special-purpose computer that manages ports, inspects, and filters network packets and determines whether to allow packets into the DMZ. Firewalls come in two varieties: *static packet filtering firewalls* that block packets based on the source and destination addresses in each packet and *stateful packet filtering firewalls* that block packets based on content, level of protocol, and history of packets. Stateful firewalls are sometimes called *dynamic filtering* firewalls.

Firewalls are not perfect. In fact, they are far from perfect, because they cannot block all malicious programs. Simply stated, a firewall is mainly used to manage ports and VPNs. They may not be adequate for detecting Trojan horses, preventing DDoS attacks, and thwarting email viruses. Therefore they should not give the administrator a false sense of security, but instead constitute the first step in establishing a TP between users and information.

A *proxy server* is a special-purpose computer that sits between a user and the enterprise server. It intercepts all requests to the real server to see if it can fulfill the requests itself. If not, it forwards the request to the enterprise server. The purpose of a proxy server is twofold: to improve security and to enhance performance. It improves performance and security by caching incoming requests on behalf of an external Web site or user. In this way unauthorized requests can be thwarted by the proxy and never reach the inner components of the TCB. The enterprise system does not expose all of its information to the outside world—only the public portions.

A proxy server can also perform the functions of a *gateway* by accessing external pages on behalf of an internal user. Each time a user requests a page from a remote Web site, the proxy server is consulted, and if the page is already inside of the DMZ, the proxy server supplies the page instead. This avoids delays and enhances security because the entire transaction is performed within the DMZ. It is not necessary to venture beyond the firewall. Gateway proxies are also used to prevent employees from viewing unauthorized web sites.

Every good TCB needs an IDS. This is a special-purpose computer that inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate that a network or system attack is underway. It uses a variety of algorithms to detect when someone is attempting to break into or compromise the DMZ. For example, it may employ *misuse detection*—the process of comparing "signatures" against a database of attack signatures to determine if an attack is underway. Or the IDS may employ *anomaly detection* by comparing the state of the network against a "normal" baseline.

An IDS can be network based or host based. A *network-based IDS* protects an entire network, whereas a *host-based IDS* protects a single computer such as a home personal computer (PC). It can also be passive or reactive. A *passive IDS* simply logs network traffic status and only signals a human operator when an unusual pattern is observed. A *reactive IDS* automatically terminates a user session or blocks network traffic from the suspected source when it detects a suspicious pattern.

A *Web server* is a computer with special software to host Web pages and Web applications. It is the component that hosts HTTP/HTTPS and delivers HTML/XML pages to users. Figure 8.2 shows how a Web server acts like a traffic cop, handing off actual processing to other computers. For example, email messages are handed off to an email server, XML messages are handed off to the XML server for parsing, database queries are handed off to a database application server, and security functions are handed off to the LDAP directory server.

Some well-known examples of Web servers:

- Apache
- MS Internet Information Server (IIS)
- Google Web server

An LDAP server is an essential part of any TCB.[6] Its function is twofold: to participate in the authentication of users

[6]LDAP—defined by the IETF—is a relatively simple protocol for updating and searching directories running over TCP/IP.

through password storage and verification and to hand out permissions—called *privileges*—to running applications.[7] Microsoft Active Directory is a commercial example of an LDAP server. It holds the usernames and passwords of all authenticated users of an enterprise system.

Access privileges are transferred among users of TCBs through a ticketing system called *X.509 certificates*. A *certificate* is a digitally signed message that transfers privileges from the sender to the recipient. Think of a certificate as a theater ticket for getting in to the show. X.509 is a recommended standard as defined by the IEEE and International Telecommunications Union (ITU).[8] Most computer users see certificates as dialog boxes that pop up in the middle of a Web browsing session. The dialog asks the user to allow the foreign request access to his or her computer. If the user agrees, the certificate transfers permission from the user to the foreign requestor.

How do users get privileges? They do so by providing a user login and password to an authentication program to verify the authenticity of the login. The authentication program obtains each user's access privileges from a (LDAP) directory that is safely stored within the DMZ. As the user moves from one application to another, each application consults the list of user privileges to determine if he or she has the necessary access rights. For example, one user may have the right to read a database record and another user may have the right to change the record. In this way the user does not have to login repeatedly to different applications, and the entire system is protected from unauthorized internal access. Certificates and access privileges form the basis of a security infrastructure called *PKI*, which is described in Section 8.1. But first, we need to understand the basics of encryption, because PKI is based on public–private key encryption.

8.4 BASICS OF ENCRYPTION

Encryption—turning plaintext messages into secret codes—is at least 4000 years old. Encryption converts *plaintext* words into *ciphertext* using a key and an encoding algorithm. The result is called a *cipher*. The reverse process—converting ciphertext into plaintext—is called *decryption*. The *key* is a special word that enables encoding. If the same key is used to encode and decode the secret message, we say the encryption is *symmetric*. If a different key is used, we say the encryption is *asymmetric*. Cryptography is the study of ciphers, keys, and encryption algorithms.[9]

TABLE 8.1 EXCLUSIVE-OR logic: Only one of the two operands can be 1 in order to produce 1. Otherwise, EXCLUSIVE-OR logic produces a 0

EXCLUSIVE-OR	B = 0	B = 1
A = 0	0	1
A = 1	1	0

During most of its history, cryptography did not change much. Find a way to translate plaintext into ciphertext and then transfer the ciphertext to a recipient, who reverses the process using the secret *key*. The cipher is symmetric, because both parties use the same key to encode and decode the secret message. Thus the key must be protected, because anyone with the key can unravel the cipher.

Perhaps the best-known symmetric cipher is the logical EXCLUSIVE-OR cipher—widely known because of its simplicity. It performs the logical EXCLUSIVE-OR operation on each bit of the binary representation of plaintext (see Table 8.1). It works bit by bit across the plaintext by taking one bit from the plaintext word, another bit from the key and writing the EXCLUSIVE-OR as the ciphertext. To reverse the process, from ciphertext to plaintext, do the same thing over again: EXCLUSIVE-OR, the key with the ciphertext.

For example, suppose the shared secret key is 1101 and the sender wants to encrypt the plaintext 1001 and send it to the receiver, who also knows the key. Encoding is done by EXCLUSIVE-ORing each bit of the message 1001 with the each corresponding bit in the key. The same process is repeated to recover the plaintext from the ciphertext.

Sender Encodes 1001 using key *1101* as follows:

$$1101\text{EXCLUSIVE}-\text{OR}\,1001 = 0100 = \text{ciphertext}.$$

Receiver decodes 0100 as follows:

$$1101\text{EXCLUSIVE}-\text{OR}\,0100 = 1001 = \text{plaintext}.$$

Keys in the EXCLUSIVE-OR cipher are limited to no more than 2^k possible values for a key with k bits. That is, the time it takes to enumerate all possible keys is proportional to 2^k. For example, a 20-bit key can have no more than 2^{20}—approximately 4 million—distinct values. This may seem like a lot, but even a key with 128 bits is not too large for a modern computer to run through in a relatively short period of time. The EXCLUSIVE-OR cipher can be cracked by simply trying every key value from zero to $2^k - 1$. But a key with 256 bits would take a computer 2^{256} units of time to

[7] Additions to version 3 of LDAP rectified many of the shortcomings of the original LDAP and allowed LDAP servers to correctly store and retrieve X.509 attributes, but searching for them was still impossible. This is because the protocol fields, that is, the X.509 attributes, are simply transferred and stored as binary blobs by LDAPv3, with the server having no knowledge about their structure and contents. "Modifying LDAP to Support X.509-basedPKIs," D.W.Chadwick, E. Ball, M.V. Sahalayev, University of Salford, Salford, M5 4WT.

[8] X.509 is actually an ITU recommendation. Nonetheless it has been widely adopted by the Internet community.

[9] A simple definition of cryptology is the study of secret messages.

crack—many times more than the time to crack a cipher with half as many bits. In other words, the strength of a cipher is *exponentially* related to the number of bits in the key. Key length determines encryption strength. Cybersecurity needs *strong encryption* and this means ciphers with large keys.[10]

8.4.1 DES

In the 1960s a team of IBM researchers designed a symmetric cipher for commercial applications they called the *Lucifer algorithm*. Lucifer was not unique, but it was destined to become the first standard encryption technique for the US federal government. Indeed, Lucifer was adopted by the NIST for use by nonmilitary customers in 1977 and revised in 1994. Simply called the Data Encryption Standard (DES), the Lucifer algorithm has been widely used by banks, insurance companies, and handheld devices such as cellular phones. It is also known as "56-bit encryption," because it uses a 56-bit key:

> In the late 1960's, IBM's chairman Tomas Watson, Jr., set up a cryptography research group at his company's Yorktown Heights research laboratory in New York. The group, led by Horst Feistel, developed a private key encryption system called "Lucifer." IBM's first customer for Lucifer was Lloyd's of London, which bought the code in 1971 to protect a cash-dispensing system that IBM had developed for the insurance conglomerate.
>
> In 1968, the National Bureau of Standards (NBS, since renamed National Institute of Standards and Technology, or NIST) began a series of studies aimed at determining the US civilian and government needs for computer security. One of the results indicated that there was a strong need for a single, interoperable standard for data encryption that could be used for both storage and transmission of unclassified data (classified stuff was still the domain of the NSA).[11]

DES uses 64 bits: 56 for data and 8 for error-checking parity. It is also called a *block cipher* because it breaks the message into 64-bit blocks, and encodes each block, separately. There are actually four variants of DES:

1. ECB = electronic codebook (standard DES algorithm)
2. CBC = cipher block chaining
3. CFB = cipher feedback
4. OFB = output feedback mode

The DES algorithm is described in more detail in Appendix D.

Unfortunately, DES was cracked in 3 days in 1998 using a special-purpose computer. In 1999 it was cracked in 22 h using 100,000 PC working together over the Internet. Today, cracking the 56-bit DES cipher is child's play for most home computers. Using longer keys, however, can strengthen DES.

8.4.2 3DES

The easiest way to make DES stronger is to make the keys longer—three times longer, in fact. A sophisticated way to make longer keys is to encrypt many times. 3DES simply applies DES three times with three keys: Key1, Key2, and Key3. This effectively increases key length threefold, from 56 to 168 bits. It also increases the difficulty of breaking the code by a factor of 2^{112}, or about 168 years of doubling in computer processing speed.[12] 3DES is strong, but somewhat cumbersome.

8.4.3 AES

Modern symmetric encryption uses the *Advanced Encryption Standard* (AES), adopted by NIST and officially standardized by the US government in 2002. It is an alternative to DES and 3DES that uses even longer keys: 128, 192, or 256-bit keys. In May 2002, NIST adopted the Rijndael (Daemen–Rijmen) algorithm as the basis of AES [1]. A 256-bit Rijndael cipher is 2^{200} times stronger than DES and 2^{88} times stronger than 3DES. In other words, it will take 120 years of progress in computing to achieve the necessary speeds to crack AES the way that DES was cracked in 1999.

One major advantage of AES, in addition to its strength, is that Rijndael works on small machines, which means AES is suitable for industrial control applications. But it takes 10, 12, or 14 rounds, depending on key size, to encode and then to decode messages. This is slower that other symmetric codes, but not too much of a burden for modern processors—even the commodity processors used in most industrial controls. The future of symmetric encryption is AES.

8.5 ASYMMETRIC ENCRYPTION

Code breakers have learned to use exhaustive brute-force methods to defeat symmetric key ciphers. In fact, high-powered computers have become very good at finding keys for symmetric encryption. While the future of symmetric encryption is AES, even AES has faults. One of these is the exposure that comes with sharing the secret key among many users. In symmetric encryption, both the sender and receiver use the same key, which exposes the most critical piece of the cipher—the key.

When consumers use a credit card to purchase a product via the Internet, from an unknown and untrusted e-commerce

[10]The strength of a cipher is measured by how long it takes for a computer to break it. Today, strong encryption means a computer the size of the universe would need all of recorded time to break the cipher.

[11]http://library.thinkquest.org/27158/concept2_1.html

[12]Moore's law says processing speed doubles every 1.5 years. So, $1.5 \times 112 = 168$.

site, how is it possible to securely exchange a symmetric key? The act of sharing exposes both parties to a man-in-the-middle attack, because transmitting the key over the Internet is no safer that transmitting credit card numbers. And if the shared key must be kept secret, how does the consumer share the key without encrypting it?

The key distribution problem described above was solved by splitting the key into two keys—one for encrypting and another for decrypting. The idea is simple, brilliant, and difficult to implement. But if only the consumer knows the decryption key, only he or she is able to decrypt messages sent to him or her. And if everyone knows the encryption key, everyone can encrypt messages specifically for one consumer only. This is called asymmetric encryption because the public key is different than the private key, but the two go together.

In 1976 two clever mathematicians named Diffie and Hellman closed the cryptographic loophole left open by key sharing [2].[13] They found an elegant way for two parties to share a secret message without sharing the same secret key. Instead, each party shares a *public key*, but conceals his or her own *private key*. Public keys are used to encode and private keys used to decode the secret message. In this way, two parties can keep both the message and their private keys a secret. PKI is *asymmetric*, because a different key is used to encode than to decode.

Here is how it works: consumer Alice creates a private key for her personal use only. This may be done by scrambling a password, biometric data, or generating a random number. The private key is used to generate a public key, which Alice shares with everyone who wants to send her secret messages. If an e-commerce site uses her public key to encode a message to Alice, she uses her private key to decrypt it. Alice never shares her private key with anyone.

The Diffie–Hellman invention of asymmetric cryptography using a public key was profound and disconcerting to intelligence agencies of the US federal government, because it allowed anyone to build ciphers that nobody could crack—not even the powerful National Security Agency (NSA). Before the Diffie–Hellman method of key distribution was invented, the NSA routinely cracked ciphers. Afterward, criminals and law-abiding citizens alike were able to keep their messages completely secure and completely unbreakable by governments.

The history of cryptology is long and colorful—too long and colorful to do it justice in this chapter. But one of the most interesting events of recent history is the peculiar case of Phil Zimmermann and the US Customs. Zimmerman was accused of trafficking in munitions simply because he wrote a computer program that implemented the Diffie–Hellman distribution algorithm. In 1995 Charles Gimon described the essence of asymmetric encryption and Phil Zimmermann's program called *Pretty Good Privacy* (PGP):

> In 1976, a completely new way to encrypt messages was published by Whitfield Diffie and Martin Hellman. This new method was called *public key encryption*. In this system, each person has two keys, a public key and a private key. The public key is broadcast around so anyone can use it, the private key is known only to the owner. You can encode a message with the recipient's public key so that only they can decode it with their private key. This public key encryption not only provides privacy, it also makes it possible to be certain that only the sender wrote the secret message you received. It ensures both privacy and identity.
>
> Public key encryption is fantastically difficult for even computers to break. The longer you make the keys, the more difficult public key encryption is to break. You can make the keys long enough so that, using today's technology, anyone's best guess is that it would take so-and-so many billions of years to break the code. One cute phrase you hear to describe this situation is "acres of Crays."[14] There's even wild talk of making keys so long that using the code breaking methods we have right now, you'd need a computer with more circuits than there are atomic particles in the known universe working for a longer period of time than has passed since the Big Bang to break it. In other words, a metaphysically unbreakable code—talk about tough math homework.
>
> Many companies, including AT&T, SCO and Sun Microsystems, have used public key encryption in their products. In order to give the power of public key encryption to folks like you and me, a programmer in Boulder, Colorado named Phil Zimmermann put together a shareware program called PGP—"Pretty Good Privacy"—which lets anyone with a PC use public key cryptography.
>
> Governments like ours have a healthy respect for cryptography; it's sometimes said that the U.S. and Britain won the Second World War by breaking German and Japanese codes. In the United States, strong, "unbreakable" encryption is considered a weapon for export purposes, just like hand grenades or fighter planes are. In theory, it's illegal to export public key cryptography, on paper or as a computer program.
>
> In 1991, right after the Gulf War, there was a bill before the U.S. Senate (S.266) that would have had the effect of banning public key encryption altogether. Faced with this situation, some activists in the [San Francisco] Bay Area decided that if they could spread public key encryption around widely enough, the genie would be out of the bottle and there'd be no way for Uncle Sam to get it back in again. They took a copy of Zimmermann's program and uploaded it to as many bulletin boards and Internet sites as they could.
>
> It took the Feds two years to react. In February 1993, Mr. Zimmermann received a visit from Customs. Even though he didn't do the uploading himself, the Feds say that

[13]Actually, three British Security Service researchers—Ellis, Cocks, and Williamson—discovered public key encryption in 1968/1969, but because their work was classified, they could not publish their results. Diffie and Hellman discovered public key encryption, independent of the British Security Service researchers.

[14]At one time, Cray computers were the fastest computers on the planet.

> Zimmermann allowed his program to be uploaded to Internet sites that can be reached from anywhere in the world, and therefore he has supposedly exported a munition without a license. It sounds like something an oily guy in Miami or Beirut would be involved in—but a computer geek in Boulder, Colorado?
>
> David and Goliath aspects aside, the case is important for two reasons. The obvious one is the First Amendment one—computer software ought to be considered speech, something that Congress isn't supposed to pass any law abridging the freedom of. Anyway, encryption is just math, and restricting or banning it isn't that much different than banning the knowledge that two plus two equals four.
>
> The other thing about the [Zimmermann incident] is its impact on America's software industry. Restricting the export of strong encryption is a joke—you can buy it shrink-wrapped in Moscow. The restrictions are an outdated, artificial leg-iron on American companies, and if they were enforced on everybody, it would make American encryption software a second-rate choice in every other part of the world. Public key encryption lets you do secure transactions on the Internet. That means buying and selling and free enterprise—all the things that we won the Cold War for—with little risk of theft or fraud. It's a shame that exporting what could be a great crime-fighting device could end up being a crime itself. [3]

In 2016 the question of strong encryption versus law enforcement came to a head with the Apple versus FBI conflict. Federal judge Sheri Pym asked Apple to help the FBI unlock an iPhone belonging to Syed Farook, the terrorist responsible for killing 14 and wounding 22 people in shootings in San Bernardino, California. The request to provide "reasonable technical assistance" to the US authorities was declined by Apple.[15]

The FBI dropped the case a day before it was to appear in court. Apparently, the government had found a way to break into the iPhone without Apple's help. By dropping the case, and several similar cases since, the government has sidestepped a ruling that may prevent them from every breaking into encrypted smartphones. The idea of a backdoor has since been abandoned, but not fully decided. A backdoor would allow the government to bypass encryption, but also weaken security. Encryption is either strong or it is not, depending on whether it can be bypassed in some way.

The sociopolitical implications of encryption mathematics are obvious from the description above. Encryption is critical to secure operation of IT infrastructure. But it is not just a topic for computer experts because it affects everyone. The remainder of this chapter will be devoted to the discussion of encryption's role in establishing a PKI. A thorough discussion of the politics of encryption will be left to another author.

[15]https://en.wikipedia.org/wiki/FBI%E2%80%93Apple_encryption_dispute

8.5.1 Public Key Encryption

The nontechnical reader may want to skip the following section, which describes, by example, how PKI works. It is the backbone of trusted computing. Without PKI, privacy would not be possible in the Internet Age. However, it is also a highly mathematical topic.

The Diffie–Hellman paper describes the concept of PKI, but it did not describe how to actually do it. The problem was that translation from plaintext to ciphertext had to be one way. That is, the process had to be irreversible. Otherwise, the receiver of a secret message could work backward and discover the sender's key. Most mathematical operations are two way: $3+2=5$ is reversible to $3=5-2$, and division, $6/3=2$, can be reversed by multiplication $2\times 3=6$. Asymmetric encryption needed a mathematical operation that worked one way but not the other way.

In 1977 three mathematicians—Ronald Rivest, Adi Shamir, and Leonard Adleman (RSA)—started their journey into the annals of encryption history by attempting to prove Diffie and Hellman wrong. Instead, they showed how to implement the Diffie–Hellman idea, which led to the famous RSA cipher in 1977 [4]. Today, RSA is the most common form of encryption used in PKI. Appendix E shows how to do PKI using RSA.

PKI is an extremely clever application of big numbers—really big numbers![16] RSA is based on prime numbers raised to large powers, which results in extremely large numbers.[17] The numbers are so large that it takes a computer to add, subtract, multiply, and divide them. In fact, the larger the number, the better, because a code breaker must be able to find extremely large prime numbers just to start the process of cracking an RSA cipher. Prime numbers containing hundreds of digits are common, and primes with millions of digits are well known to the intelligence community.

RSA translates a series of plaintext words into a series of codewords that look random (see Fig. 8.3). This makes it difficult for code breakers to analyze long sequences of codewords using pattern-matching software to unravel the key. Instead of producing an intelligible pattern, pattern analysis produces random noise.

Appendix E illustrates the RSA technique using December 7, 1941 (12/7/41), as an example of a message to be sent from Honolulu to Washington, DC. Honolulu uses Washington's public key to encode, and Washington uses its own private key to decode. So 12/7/41 is encoded as {23, 13, 6} using Washington's public key P = (55,3). When Washington receives the ciphertext, it decodes {23, 13, 6} into (12, 7, 41) using its private key, V = (55, 27).

[16]Public key encryption uses numbers in excess of 200 digits long!

[17]A prime number is a positive number that is divisible by one and itself only. Prime numbers can be found by computerized mathematical sieve techniques that take time proportional to the size of the prime number.

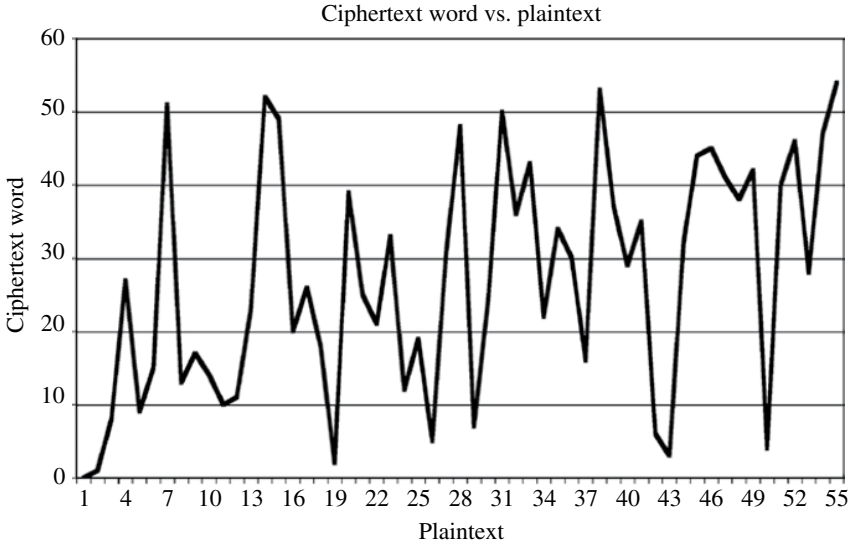


FIGURE 8.3 RSA encryption produces a seemingly random stream of codewords from plaintext. The codewords were produced from the public key P = (55, 3).

Honolulu does not know V, and so only Washington can decode the message. However, there are other private keys that can also decode the message. For example, V = (55, 67) also unscrambles the cipher. But nobody knows the exact values used in these other keys. Why?

PKI cleverly uses the one-way property of *modulo arithmetic*. Its strength is based on the (large) size of keys, which are large prime numbers. While these are not difficult to compute, there are so many of them with hundreds of digits that it takes a long time to crack.

8.5.2 RSA Illustrated

The RSA algorithm is the foundation of PKI. It is maintained by a network of CAs, described in Appendix E, and is operationally very simple. First, a consumer Alice creates a private key only she knows. Then, she creates a public key that everyone knows, based on the same pair of prime numbers. This pair must be very large because the RSA encryption becomes unraveled once these primes are discovered.

The following illustrates the RSA encryption as shown in Figure 8.4. Given two prime numbers, p and q, and a plaintext message such as "Now is the time," RSA calculates a public and private key, encodes the plaintext message using the public key, and then decodes the cipher using the private key. Keyboard characters are converted into numerical equivalents, numerically processed, and then converted back into alphanumeric characters as needed.

Figure 8.4 displays the numerical value of codewords (encrypted plaintext characters) versus the alphanumeric plaintext characters "0" through "9," "A" through "Z," and "a" through "z" using the RSA algorithm described in Appendix E. Note that each public key (p, q, e) produces a different graphical display, but they all look random. Also note that each private key (p, q, d) produces the *same* result—they all correctly decode the cipher. Only the public key determines the cipher. The private key simply decodes the cipher, returning it back to its original plaintext.

8.5.3 Shor's Algorithm

PKI is strong encryption because it separates keys into public and private. Its strength is based on secret prime numbers p and q. Brute-force methods of discovering p and q have been attempted using powerful computers for decades. As computers become more powerful, cryptographers increase the size of p and q, making brute-force methods impractical. However, quantum computing is about to change this, because quantum computers can solve Shor's algorithm in a matter of seconds. The solution to Shor's algorithm is the two prime numbers p and q!

$$\text{Shor's equation}: \left(N - pq\right) = 0$$

Peter Shor proposed his famous factorization equation in 1994 for the purpose of breaking the RSA code. For example, $(55 - 5(11)) = 0$ is a solution to Shor's equation. But without knowing in advance what p or q are, the task of factoring $N = 55$ into its prime factors may take years of computer time, unless the computer can perform many comparisons simultaneously. As it turns out, quantum computers are very good at performing simultaneous calculations.

Quantum code breaking solves Shor's equation by computing all possible combinations of p and q at once. Instead

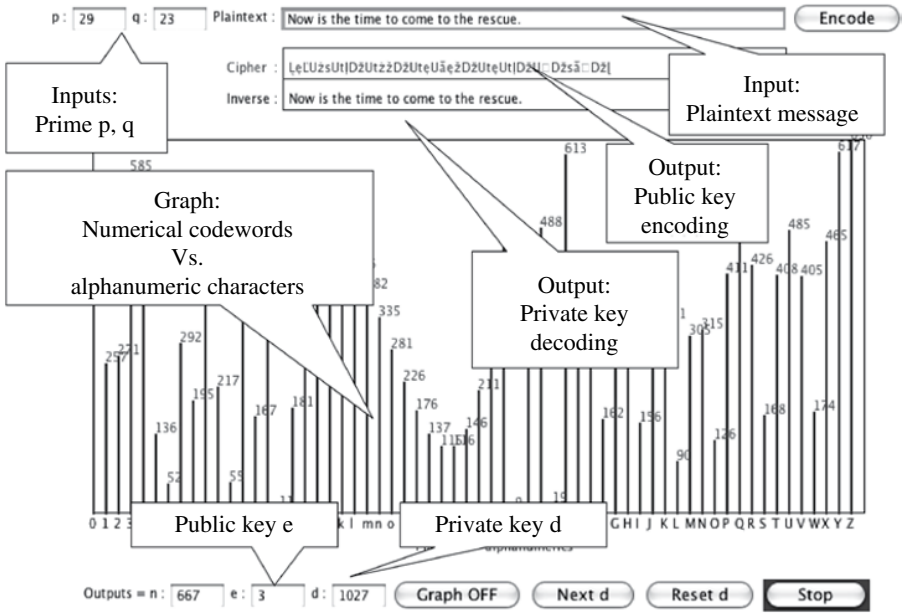


FIGURE 8.4 Screen display showing "randomized" ciphertext output from the RSA algorithm.

of testing each pair (p and q) one after the other, a properly set up quantum computer computes all possible pairs and discards the pair that does not equate ($N - pq$) to zero. The first quantum computer to solve Shor's equation was demonstrated in 2001.[18] A research group at IBM factored 15 into $p = 3$ and $q = 5$. By 2014, the largest equation solved was $(56{,}153 - (233)(241))$. Eventually, a quantum computer will be able to factor very large numbers. This will make it practical to crack the RSA algorithm.

The RSA algorithm's days are numbered because quantum computers are on the horizon. When they are able to solve Shor's algorithm in a matter of seconds, PKI will no longer guarantee security. However, the same quantum computing technology that enables code breaking will be used to prevent it. Quantum key distribution (QKD) will replace the RSA algorithm because QKD does not depend on primes.

8.6 PKI

The following is a superficial introduction to the vast and complex topic of PKI. Some accuracy will be sacrificed for simplicity. The foundation of TP in cybersecurity is PKI, which is only as good as the public and private keys used to encrypt and decrypt messages. Thus, secure key management becomes critical. Hackers and crackers will try to break into a system and steal passwords, for example. If the attacker unravels the encryption key, all passwords will be exposed. Cracked password files give attackers access to bank accounts and critical databases, for example.

PKI combines encryption, key management, and user authentication into a comprehensive system for implementing TP and TCB. It enables users who do not know each other, and perhaps may never meet in reality, to trust one another in cyberspace. PKI defines the way users exchange mutual trust regardless of their location in the global Internet.

PKI has to manage authentication, privileges, keys, and secrecy. In addition, PKI has to be standardized so that authentication, privileges, keys, and secrecy can be exchanged among different systems. Standardization is a critical element of PKI.

Two IETF working groups—PKIX (PKI X.509) and SPKI (Simple PKI)—continue the process of developing PKI standards. Some of the more important RFCs related to PKI are:

RFC 2401 (Security Architecture for the Internet Protocol, November 1998)

RFC 2437 (PKCS #1: RSA Cryptography Specifications Version 2.0, October 1998)

RFC 2527 (Internet X.509 Public Key Infrastructure Certificate Policy and Certification Practices Framework, March 1999)

RFC 2692 (SPKI Requirements, September 1999)

RFC 2693 (SPKI Certificate Theory, September 1999)

RFC 2898 (PKCS #5: Password-Based Cryptography Specification, Version 2.0, September 2000)

[18]https://en.wikipedia.org/wiki/Shor%27s_algorithm

8.6.1 Definition of PKI

PKI has been defined in a number of ways, but the following definition was selected because of its simplicity:

> A public-key infrastructure (PKI) is a full system for creating and managing public keys used for encrypting data and exchanging those keys among users. A PKI is a complete system for managing keys that includes policies and working procedures. PKI is about distributing keys in a secure way. Whitfield Diffie and Martin Hellman developed the concept of asymmetric public-key cryptography in 1976, but it was RSA (Rivest, Shamir, Adleman) Data Systems that turned it into a workable and commercial system. Today, RSA is the most popular public-key scheme.[19]

The PKI system described above includes the management of certificates (permissions or privileges shared by a sender and receiver of documents), the management of encryption, and the management of authentication. Therefore, it is more comprehensive than simple encryption or simple authentication. When combined with XML, PKI is called XKI, and usually incorporates login and password authentication services as well as certificate services.[20]

The goals of PKI are:

- *Authentication*: PKI assures that the sender is whom he or she claims. This is done by a combination of PKI and the use of certificates.
- *Integrity*: PKI guarantees the integrity of the message, for example, that an intermediary has not modified the message during transit.
- *Confidentiality*: PKI assures the message is decodable only by the intended recipient. Encrypting the message and authenticating the recipient guarantees confidentiality.
- *Non-repudiation*: PKI assures that verified parties signed the message. This is implemented in PKI by authenticating the users and trusting the CAs. Therefore, PKI guarantees that both parties cannot disavow or deny involvement with the transaction. This is achieved by attaching the private keys of the sender (receiver) to the message.

In the following, we illustrate how a typical PKI system works and how each goal above is met by PKI.

8.6.2 Certificates

Certificates are tickets for communicating trust. Like a passport or birth certificate, X.509 certificates have become the ad hoc standard for exchanging trust over the Internet. The assumption underlying certificates is that they emanate from a trusted source—the so-called certificate authorities. Ultimate trust is based on a root authority that says you are who you say you are and that you have the privileges stated on your certificate.

X.509 certificates are created by a CA, that is, a trusted LDAP server or trusted third party. Minimally, they contain the identities and keys of the parties that want to enter into a trusted relationship. For example, the following certificate contains the identity and public keys of Alice and Bob, two users who want to enter into a trusted relationship:

Real name	Username	Public key
Alice	A	Public (3, 3233)
Bob	B	Public (17, 6345)

For simplicity, assume Alice's username is A and Bob's username is B. The certificate contains other information, but this simple example will be sufficient to illustrate how PKI works.

Figure 8.5 illustrates how PKI works. Alice, working within her TCB, sends a message to Bob, working within his TCB, through a TP established by encryption, authentication, and certificates. The process is started by Alice, who obtains Bob's public key from her CA, signs the message labeled Msg_to_Bob, encrypts it, and sends it to Bob. The CA can be an LDAP server where certificates like the one above are stored and served up whenever the TCB needs to transfer trust from one user to another.

Next, Alice digitally signs her message and encodes her signature using her private key, so only she can unlock her signature. This guarantees that only Alice has signed the message. She then appends her public key to the encoded message so that Bob can verify that she sent the message. A *digital signature* is an electronic code that authenticates the identity of the sender of a message or the signer of a document and possibly to ensure that the original content of the message or document is unchanged. By using her private key, Alice can determine that the message was actually created and sent by her, because only she knows her private key.

In the final step Alice uses Bob's public key to encrypt Msg_to_Bob and then sends the encrypted and signed message to Bob. This message contains the original message content, Alice's digital signature and Alice's public key, all encrypted using Bob's public key. By using Bob's public key to encrypt the whole package, only Bob can decode it, and therefore, only Bob and Alice know what was sent. This assures privacy. By including Alice's digital signature, only Alice could have sent the message, because only Alice knows her private key. And by including Alice's public key, Bob can verify that Alice is the sender. Bob knows the message came from Alice because she is an authenticated user.

[19]http://www.linktionary.com/p/pki.html

[20]Defining boundaries between LDAP directories, authentication, and PKI is a moving target because these technologies appear to be merging into a comprehensive system of security.

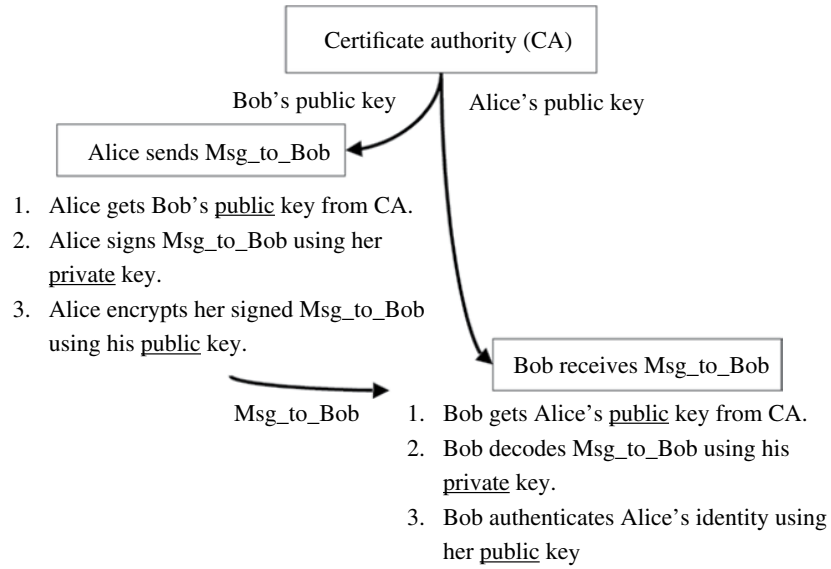


FIGURE 8.5 An Example of PKI: Alice sends Msg_to_Bob to Bob. Issuing certificates and encrypting the message using the RSA algorithm establishes a trusted path.

Alice knows she is the author of the message because her private key is embedded, and both know the message has not been tampered with because it is encrypted.

At the other end, Bob receives the encrypted Msg_to_Bob from Alice. He looks up her identity from the CA, and it returns a certificate containing Alice's public key. Bob uses his private key to decode the message, which also contains Alice's public key. He verifies that this message actually came from Alice by comparing the public key obtained from the CA with the public key decoded from Msg_to_Bob. If Alice tries to repudiate that she sent the message, she will have a difficult time, because her private key was used to encrypt the signature. Only she could have done this.

The only way someone besides Bob and Alice could have sent or received the message is if someone stole their private keys. Certificates guarantee that Bob and Alice are who they say they are; RSA encryption guarantees security; privacy and enforceability is assured by certificates and CAs; and non-repudiation is assured by digital signatures.

PKI establishes TP. If implemented correctly, PKI assures a TCB. But if the CA is cracked, PKI cannot guarantee security, privacy, enforceability, or non-repudiation. Therefore, it is critically important that keys be protected, CAs be secure, and the RSA algorithm never be cracked.

CAs are hierarchical directories that vouch for one another. At the highest level within the hierarchy—the root CA—the root CA signs certificates itself. That is, the root CA vouches for itself. The top-level CAs use digital signatures and certificates to vouch for sublevel CAs. Thus, trust is passed down from a root CA to sublevel CAs. Certificates are signed by trusted CAs using the private key and authenticated using the recipient's public key, just like any other message. Remember, public keys are used to encrypt, and private keys to decrypt. Thus a CA encrypts each certificate using the user's public key. The user decrypts the certificate using his or her private key. The certificate can be verified just like any other message.

8.6.3 Blockchain

An alternative to the CA is the blockchain invented by mysterious Satoshi Nakamoto in 2009 as infrastructure for bitcoin. See Chapter 17 for more on bitcoin. A blockchain is a distributed ledger containing blocks of transactions secured by public–private keys. Each block is linked to the next block going forward in time, so the ledger can only add new information at the end. The chain is a ratchet, which means it can only be updated in order, from oldest to newest transaction.

Blockchain security differs from PKI security based on CAs because the blockchain is distributed while the CA is centralized. The CA system may be hierarchical, but there is only one CA at each level in the hierarchy. This means consumers must trust the central authority. Stuxnet used a stolen black market certificate to authenticate itself, so we know that RSA PKI is fallible.

Blockchain distributes authority to equals called peers. The blockchain is copied to peers called nodes in a peer-to-peer (p2p) network. Instead of ironclad trust assigned to the top of the CA hierarchy, blockchain trust is distributed across the p2p network. Consensus of opinion is derived by sharing responsibility for updates across the p2p network. A transaction is added to the chain only after consensus grants it.

Consensus is defined in different ways in different implementations of blockchains. It is achieved by a proof-of-work (PoW) algorithm in original bitcoin. The first node to find a *nonce* (number used once) sends it to the remaining nodes to verify. Consensus is achieved by a majority of nodes testing and verifying the first node's claim. A nonce is a valid cryptographic key that is also smaller than a threshold set by the p2p network. Finding the nonce is difficult and somewhat random. It is also a computationally intense and expensive way to obtain consensus, so other methods have been proposed.

Proof of stake (PoS) is an alternative to PoW first deployed by Ethereum. It operates more like a pure democracy. Each node has a vote proportional to its stake in the chain. This is often determined financially, but it can also be determined hierarchically. For example, the president of a company may have a larger stake than other employees, or law enforcement may have more stake than citizens. Unfortunately, PoW and PoS can be circumvented by cartel-like behavior where a collection of nodes conspires to outvote or outcompute everyone else.

Blockchain security is an alternative to the CA hierarchy that distributes trust to a network of users and consumers. It places trust in equals, rather than a CA. However, it is subject to Gause's competitive exclusion principle that rewards preferential attachment. The benefits of a p2p network are diminished by preferential attachment.

8.6.4 FIDO and WebAuth

FIDO is an industry standard protocol for securely authenticating online users through a "lightweight certificate" system that prevents man-in-the-middle attacks and eliminates the need for passwords. At the time this was written, FIDO2 was adopted by the W3C and called *WebAuth*. It is the standard authentication protocol for major Web sites and cloud computer systems.

FIDO/WebAuth operates much like the CAs and TLS, but it is more general, supporting physical keys, 2FA, and biometrics. It replaces the simple password method of user authentication by shifting identity to a device rather than a person. Of course, if a person owns the device, authentication still identifies the person as an authorized user. Here is how it works:

- When a user connects to a Web site such as Google.com or Facebook.com, the Web site immediately *challenges* the connection by downloading a random number to the user. This number is used to seed the production of a private–public key pair.
- The user generates a unique public and private key pair. The public key is sent to the Web site to be used to encrypt messages to the user. The private key, of course, is kept private and typically protected on the user's device through an additional layer such as a fingerprint, voiceprint, or face print.
- The Web site encrypts messages to the user using the user's public key. The user uses his or her private key to decrypt and to digitally sign messages to the Web server. This continues for as long as the session continues.
- The process is repeated for each session. Thus, even if encryption is compromised, it is only valid for as long as the session.

Like the CA system, FIDO is only as secure as the public and private key pair. A user's private key must be secured, typically by encrypting it with a symmetric algorithm so it can only be opened by a user's PIN, password, and so on. Most smartphone devices, for example, contain a cryptographic chip that secures passwords, fingerprints, voiceprints, and others unique to the device. The device becomes part of the authentication process because it contains cryptographic data.

8.6.5 Mathematics of Passwords

It is highly likely that tokens, WebAuth, or biometric "fingerprints" such as face and voice recognition will replace passwords as a method of authentication. But passwords may continue to be used in certain situations and in legacy systems. Therefore, it is important to understand the mathematics of passwords and how IT professionals can make password authentication more secure.

Picking a common password is perhaps the most often cited mistakes consumers make. Hackers try these, first, in what is called a dictionary attack [5]. Over a half million passwords like this have been used in dictionary attacks to break into consumer accounts. A *pwned* password is any password known to have been successfully exploited. The word *pwned* is shorthand for *pounded*, a term used by video gamers to denote extreme defeat or "taking a pounding." Before setting up a new password, users should consult Web site https://haveibeenpwned.com/Passwords to test whether the password has been *pwned*:

1. 123456
2. password
3. 12345678
4. qwerty
5. 12345
6. 123456789
7. letmein
8. 1234567
9. football
10. iloveyou
11. admin

12. welcome
13. monkey
14. login
15. abc123
16. starwars
17. 123123
18. dragon
19. passw0rd
20. master
21. hello
22. freedom
23. whatever
24. qazwsx
25. trustno1

A well-designed Web site and IT server will never store user passwords directly. Instead, passwords should be randomized by a so-called hashing function that scrambles passwords, making them look like noise. A hash function is a type of symmetric cipher that uses a secret key to scramble each password before it is stored in the server. The server should never store usernames and passwords in plaintext. Instead, passwords should be hashed into a unique sequence using a one-way hash function. One-way hashing means that the original password cannot be reconstructed by reversing the hash function.

Here is a simple one-way hash function on characters converted into numbers, say, in the interval [1, 72]: N^3 mod 71. The first 10 numbers in [1, 72] are scrambled as follows.

Number	Hashed number
1	1
2	8
3	27
4	64
5	54
6	3
7	59
8	15
9	19
10	6

Criminals have developed extremely clever algorithms for narrowing down the search for passwords, even when hash functions are used. The best defense is to use long passwords containing numerals and characters that make little sense relative to a dictionary or natural language. There are 72 uppercase and lowercase letters and numerals in the English language. A password and its hash containing 10 characters represent 72^{10} possible combinations. As the length of your password increases, the number of combinations a hacker must try increases exponentially.

TABLE 8.2 Sample countermeasures to vulnerabilities typical found in enterprise systems

Vulnerability	Countermeasure
Power failure	Install backup power supply
Telecom failure	Buy redundant telecom service
SYN attack	Install IDS
	Install firewall: filter ports
No IDS	Install IDS
Break-in	Install IDS
	Install firewall: filter ports
	Install latest patches
Clear password file	Encrypt password files
No backup	Do periodic backups
No firewall filter	Install firewall: filter ports
No antiviral SW on desktop	Install patches
	Install antivirus SW
Clear XML/HTML	Install HTTPS/SSL
	Install PKI/VPN
Clear browser use	Time-out inactive sessions
Password not changed	Change password periodically
War dialing	Close modem ports
No HTTPS/SSL	Install HTTPS/SSL
Browser session open	Time-out inactive sessions
Weak encryption	Install 3DES or AES
	Install PKI
Weak LDAP in applications	Install LDAP directory
	Modify applications
Buffer overflow	Install patches
	Update patches
Weak OS patches	Update patches
	Install IDS
	Install firewall: filter ports
Open Wi-Fi ports	Install IDS
	Install firewall: filter ports
	Encrypt Wi-Fi sessions
	Authenticate Wi-Fi users
Open modem	Close dialup modems or use VPN
Open FTP ports	Close FTP or filter ports
Firewall filter off	Turn on firewall filtering

8.7 COUNTERMEASURES

What other countermeasures should an enterprise system use to assure cybersecurity? Table 8.2 contains a list of typical countermeasures for the vulnerabilities described in this and the previous chapters. In general, countermeasures consist of:

- Providing backup to power and telecommunications services.
- Installing and operating at least one IDS.
- Installing and operating at least one firewall.
- Installing and updating vendor-released software patches.
- Encrypting and hashing password files and periodically updating passwords.

- Perform frequent backups.
- Manage ports, especially dial-up modem ports.
- Use symmetric and asymmetric encryption to achieve desired level of security.
- Security can be achieved in layers: HTTPS/SSL at the low end and full PKI at the high end. 3DES/AES can be used where appropriate.

Cybersecurity is a trade-off between expense, effort, inconvenience, and privacy, security, and target hardening. Strong encryption protects the Internet from attack, but it also protects the terrorist and hacker. Surveillance infringes on privacy, but it is also a weapon in the global war on terrorism. High-assurance systems may be secure, but users are inconvenienced and productivity suffers. Cybersecurity is a balancing act.

Richard Pethia, Director of CERT, gave the following testimony before the subcommittee of the US House in 2003:

> The current state of Internet security is cause for concern. Vulnerabilities associated with the Internet put users at risk. Security measures that were appropriate for mainframe computers and small, well-defined networks inside an organization are not effective for the Internet, a complex, dynamic world of interconnected networks with no clear boundaries and no central control. Security issues are often not well understood and are rarely given high priority by many software developers, vendors, network managers, or consumers. [6][21]

Pethia goes on to list the following general vulnerabilities of cyberspace:

- Other critical infrastructures are becoming increasingly dependent on the Internet and are vulnerable to Internet based attacks.
- Cyberspace and physical space are becoming one. The growing links between cyberspace and physical space are being exploited by individuals bent on causing massive disruption and physical damage.
- System administration and management is often being performed by people who do not have the training, skill, resources, or interest needed to operate their systems securely.
- Users often lack adequate knowledge about their network and security. Thus misconfigured or outdated operating systems, mail programs, and Web sites result in vulnerabilities that intruders can exploit. A single naive user with an easy-to-guess password can put an entire organization at risk.
- Product security is not getting better: developers are not devoting sufficient effort to apply lessons learned about the sources of vulnerabilities. In 1995 CERT received an average of 35 new reports each quarter, 140 for the year. By 2002, the number of annual reports received had skyrocketed to over 4000. Vendors concentrate on time to market, often minimizing that time by placing a low priority on security features.
- It is often difficult to configure and operate many products securely.
- There is increased reliance on "silver bullet" solutions, such as firewalls and encryption, lulling organizations into a false sense of security. The security situation must be constantly monitored as technology changes and new exploitation techniques are discovered.
- Compared with other critical infrastructures, the Internet seems to be a virtual breeding ground for attackers. Unfortunately, Internet attacks in general, and denial-of-service attacks in particular, remain easy to accomplish, hard to trace, and a low risk to the attacker. Technically competent intruders duplicate and share their programs and information at little cost, thus enabling novice intruders to do the same damage as the experts. In addition to being easy and cheap, Internet attacks can be quick. In a matter of seconds, intruders can break into a system; hide evidence of the break-in; install their programs, leaving a "backdoor" so they can easily return to the now compromised system; and begin launching attacks at other sites.
- Attackers can lie about their identity and location on the network. Senders provide their return address, but they can lie about it. Most of the Internet is designed merely to forward packets one step closer to their destination with no attempt to make a record of their source. There is not even a "postmark" to indicate generally where a packet originated. It requires close cooperation among sites and up-to-date equipment to trace malicious packets during an attack. Moreover, the Internet is designed to allow packets to flow easily across geographical, administrative, and political boundaries. Consequently, cooperation in tracing a single attack may involve multiple organizations and jurisdictions, most of which are not directly affected by the attack and may have little incentive to invest time and resources in the effort. This means that it is easy for an adversary to use a foreign site to launch attacks at US systems. The attacker enjoys the added safety of the need for international cooperation in order to trace the attack, compounded by impediments to legal investigations. We have seen US-based attacks on US sites gain this safety by first breaking into one or more non-US sites before coming back to attack the desired target in the United States.
- There is often a lack of unambiguous or firmly enforced organizational security policies and regulations.
- There is a lack of well-defined security roles and responsibilities or enforcement of accountability in many organizations, including failure to account for

[21]Testimony given before the House Select Committee on Homeland Security Subcommittee on Cybersecurity, Science, and Research and Development, June 25, 2003. http://hsc.house.gov/files/Testimony.pethia.pdf.

security when outsourcing IT services, providing security awareness training for all levels of staff, nonexistent or weak password management, and poor physical security leading to open access to important computers and network devices.

- Other practices lead to:
  - Weak configuration management that leads to vulnerable configuration.
  - Weak authentication practices that allow attackers to masquerade as valid system users.
  - Lack of vulnerability management practices that require system administrators to quickly correct important vulnerabilities.
  - Failure to use strong encryption when transmitting sensitive information over the network.
  - Lack of monitoring and auditing practices that can detect attacker behavior before damage is done.

Finally, Pethia recommends the following remedies and actions:

- Incentives for vendors to produce higher quality IT products with security mechanisms that are better matched to the knowledge, skills, and abilities of today's system managers, administrators, and users. For example:
  - Vendors should ship their products with "out-of-the-box" configurations that have security options turned on rather than require users to turn them on.
  - The government should use its buying power to demand higher quality software. The government should consider upgrading its contracting processes to include "code integrity" clauses, clauses that hold vendors more accountable for defects in released products.
- Wider adoption of risk analysis and risk management policies and practices that help organizations identify their critical security needs, assess their operations and systems against those needs, and implement security improvements identified through the assessment process. What is often missing today is management commitment: senior management's visible endorsement of security improvement efforts and the provision of the resources needed to implement the required improvements.
- Expanded research programs that lead to fundamental advances in computer security. For example:
  - Make software virus-resistant/virus-proof.
  - Reduce implementation errors by at least two orders of magnitude.
  - Develop a unified and integrated framework for all information assurance analysis and design.
  - Invent rigorous methods to assess and manage the risks imposed by threats to information assets.
  - Develop quantitative techniques to determine the cost/benefit of risk mitigation strategies.
  - Develop methods and simulation tools to analyze cascade effects of attacks, accidents, and failures across interdependent systems.
  - Develop new technologies for resisting attacks and for recognizing and recovering from attacks, accidents, and failures.
- Increase the number of technical specialists who have the skills needed to secure large, complex systems.
- Increase awareness and understanding of cybersecurity issues, vulnerabilities, and threats by all stakeholders in cyberspace. For example, children should learn early about acceptable and unacceptable behavior when they begin using computers just as they are taught about acceptable and unacceptable behavior when they begin using libraries.

8.8 EXERCISES

1. Which of the following is NOT in the IEEE X.509 standard?
   a. Password standard
   b. Integrity of information
   c. Confidentiality of information
   d. Non-repudiation of ownership
   e. Authentication of users
2. A secure link between user and system is defined as (select only 1)?
   a. VPN
   b. PKI
   c. TCB
   d. Trusted path
   e. Certificate
3. A TCB is defined as:
   a. The Country's Best Yogurt
   b. An example of an Internet threat
   c. A mechanism for enforcing minimal security
   d. A malicious program that travels via the Internet
   e. A protocol for ensuring authentic users
4. The DMZ enforces:
   a. Enterprise computing standards
   b. PKI standards
   c. X.509 standards
   d. Complete security
   e. A security policy
5. Which of the following guarantees a secure enterprise system?
   a. Passwords
   b. Biometrics
   c. PKI
   d. X.509 certificates
   e. None of the above
6. SSL/TLS/HTTPS is a protocol for:
   a. Serving X.509 certificates to users
   b. Authenticating users

c. Encrypting communication between user and Web server
d. Encrypting credit card numbers
e. Catching man-in-the-middle thieves

7. Tunneling is a technique used in:
   a. 3DES
   b. RSA
   c. PKI
   d. DMZ
   e. VPN
8. RSA is a type of:
   a. Asymmetric encryption
   b. Authentication
   c. Password
   d. Biometric
   e. VPN
9. An IDS is a special-purpose computer (or software) for:
   a. Checking passwords
   b. Preventing break-ins
   c. Non-repudiation detection
   d. Information assurance
   e. Detecting suspicious data transmission patterns
10. DES and Triple DES evolved out of a project known as:
    a. Lucifer
    b. Hannibal
    c. Diffie–Hellman
    d. Zimmerman
    e. Rivest, Shamir, and Adleman
11. The strength of an encryption algorithm is measured by:
    a. The algorithm
    b. The AES standard
    c. The number of bits in its keys
    d. The secrecy of its algorithm
    e. FIPS compliance
12. Public keys are stored in:
    a. Personal computer address books
    b. Certificate authorities
    c. Internet DNS
    d. NSA servers
    e. Autonomous systems
13. A proxy server is often used to:
    a. Increase efficiency
    b. Decrease the cost of an enterprise system
    c. Enforce PKI security
    d. Block unauthorized users
    e. Deflect DOS attacks
14. The heart of user authentication is a server called:
    a. LDAP server
    b. Email server
    c. Certificate authority
    d. RSA encryption
    e. HTTPS/SSL
15. In the example of Alice sending a message to Bob, what mechanism ensures non-repudiation?
    a. The public key
    b. The certificate
    c. Alice's private key is in the message
    d. Bob's private key is in the message
    e. Alice's public key is in the certificate

8.9 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. How would you apply the kill chain method of malware detection and deflection to the TCB and TP? Give an example.
B. Smartphone apps often violate their user's privacy by screen scraping or even spying on uses through the camera. Propose measures for defeating apps that spy and explain what exploits each prevents.
C. Explain why longer passwords are more difficult to crack than shorter passwords. What does the length of a password have to do with breaking codes?
D. Explain in your own words how SSL/TLS works when using a browser in HTTPS mode to enter your social security number securely. Your explanation must include the CA (certificate authority), public–private key exchange, and the Social Security number to be protected.
E. Online banking allows cell phone users to send a photograph of a check to an automatic teller machine for depositing. Describe the trusted path between the cell phone and the bank's database containing the account where the money is deposited.

REFERENCES

[1] Daemen, J. and Rijmen, V. *AES Proposal: Rijndael, AES Algorithm Submission*, September 3, 1999. Available at http://www.nist.gov/CryptoToolkit. Accessed June 29, 2014.

[2] Diffie, W. and Hellman, M. New Directions in Cryptography, *IEEE Transactions on Information Theory*, 22, 6, November 1976, pp. 644–654.

[3] Gimon, C. *The Phil Zimmerman Case*, February 1995. Available at http://www.skypoint.com/members/gimonca/philzima.html. Accessed June 29, 2014.

[4] Rivest, R. L., Shamir, A., and Adleman, L. On Digital Signatures and Public Key Cryptosystems. *MIT Laboratory for Computer Science Technical Memorandum*, April 1977, pp. 82.

[5] Delahaye, J.-P. The Mathematics of (Hacking) Passwords, *Scientific American*, April 12, 2019. Available at https://www.scientificamerican.com/article/the-mathematics-of-hacking-passwords. Accessed July 23, 2019.

[6] Pethia, R. D. Cyber Security—Growing Risk from Growing Vulnerability. CERT, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, June 25, 2003, pp. 1–10.

9

HACKING SOCIAL NETWORKS

A social network is an online community linked together by an e-commerce site such as LinkedIn.com, Twitter.com, and Facebook.com. It provides Web 2.0 services whereby users are the product, rather than the e-commerce site, itself. Web 1.0 was one way, almost like television. But Web 2.0 introduced the two-way Web whereby information in the form of text, sounds, pictures, and videos flows to the cloud where it is reflected back to the social network in total or to authorized followers of individual users. Social networks are the twenty-first-century version of the community center, place of worship, auditorium, stadium, and live theater of the twentieth century. They are where virtual communities form and interact.

The basic unit of communication in a social network is the meme, defined loosely as "an idea, behavior, or style that spreads from person to person within a culture—often with the aim of conveying a particular phenomenon, theme, or meaning represented by the meme."[1] A meme may be expressed by text, image, sound, or video. It has no moral or ethical value independent of a culture or set of societal rules. It may be interpreted positively or negatively. Negative or false memes are of concern in a social network because of the power of online communities to influence users and impact society in general. Meme spreading is of concern, then, to national security because of the power of social networks to spread fake news, misinformation, and propaganda and sway political elections.

[1]https://en.wikipedia.org/wiki/Meme

The following topics and concepts are surveyed in this chapter:

- Social networks emerged from the dotcom collapse of 2001 in the form of the two-way Web 2.0, whereby consumers interact by uploading content as much or more than the e-commerce Web site downloads content to the consumer. The bidirectional technology of Web 2.0 gave individuals power over entire communities and lead to abuses such as online harassment, automated propaganda in the form of botnets, and the artful spread of misinformation by nation-states. The rise of social networks saw the corresponding rise of botnets, trolls, and fake news.
- Social networks amplify and spread social memes in the form of text, audio, and video designed to influence friends and friends of friends online. Regardless of the veracity of the tweet or post, amplified memes collect momentum as they spread, often leading to information cascades—memes that circle back on themselves and magnify as their feedback loops accelerate the speed and reach of online posts.
- When social networks are infected by a botnet, the result is a botnet under the control of a botherder whose aim is to create artificial and often misleading information cascades. The effectiveness and poser of these botnets (and human propagandists) are a function of the topological structure of the social network. Highly connected and highly influential actors have more power over the

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

spreading than less well-connected and central actors. The topology of the social network matters.

- Computational propaganda is the process of spreading misinformation for political purposes. It works by combining psychological profiling of human users with digital technologies such as botnets and leveraged network topology. Adjacent online friends and tightly linked friends of friends that agree with the meme amplify belief in a meme. Groupthink induces people to believe what others claim even when it the meme is obviously false.
- A filter bubble is formed by rewarding consumers in a feedback loop of posts and news items that increase endorphins of the human brain. Filter bubbles ultimately rewire a consumer's brain via the plastic brain hypothesis that says one-sided information leads to growth of new pathways within the brain. Thus, social networks have to power to change the way consumers think.
- The outcome of filter bubble feedback is an echo chamber whereby social network consumers tend to flock together with people that believe the same as they do. This lack of open-mindedness further contributes to altering the plastic brain. The more we engage with the echo chamber, the more we accept its memes regardless of facts.
- One of the tools of meme spreading and groupthink is a form of artificial intelligence called deep learning. Deep learning is an algorithm that simulates the neural network of animal brains. It enables computers to sift through and analyze extremely large data sets called big data to identify patterns and trends. Generally, deep learning has been used to classify digital objects such as photographs—using metadata—and recognize speech from spoken words.
- The most successful deep learning algorithms thus far have been convolutional neural networks (CNNs). CNNs have been successfully used for a wide variety of big data processing tasks such as image recognition and classification, speech recognition, recommendation systems such as what a specific consumer might want to purchase next, and finding patterns in massive amounts of unstructured data. Social networks produce billions of digital objects about users that can be harvested by CNNs to better target consumers. CNNs are also capable of violating consumer's privacy by indirectly deducting user's private information such as bank accounts and passwords.
- Data brokers collect, combine, and analyze user data to infer individual's innermost life, including where you live, what you buy, how much money you have, who your friends are, and where you shop. Additionally, deep learning techniques allow big data collectors to infer passwords and other private information obtained by how you use your smartphone.
- The General Data Protection Regulation (GDPR) laws from the European Union (EU) portend to alter the social media landscape by requiring social networks to gain permission from users for the use of their data and give them the right to be forgotten, for example, erase all data collected about the consumer.
- The GDPR does not far enough, however, because it does not restrict the use of bots, harvesting cookies, and the spread of fake news. Thus, more is needed to protect consumers. Some regions of the globe are beginning to address information leaks and information cascades that harm consumers. California, for example, passed a law in 2019 requiring bots to identify themselves as bots. Other regulations are likely to follow due to congressional oversight of social networks such as Twitter.com and Facebook.com.
- The Hodges Fragility Conceptual Framework adapted for online social network resilience scores connectedness on the basis of leadership, trust, and partnerships; stability on the basis of cohesiveness, change rate, and credibility; and sustainability on the basis of intervals between exploits, recoverability, and cost of prevention.
- As an illustration only, a hypothetical social network representing the state of networks such as Twitter.com, Facebook.com, and Instagram.com circa 2018 scores a medium to low resilience rating on a Hodges framework adapted to social network resilience.
- Warnings against surveillance capitalism began to appear in 2019 following revelations regarding the surveillance and collection of private information on consumers by social networks. These warnings lead to unknown legislative actions going forward. The future of social networks is likely to be a flurry of twenty-first-century regulations.

9.1 WEB 2.0 AND THE SOCIAL NETWORK

Social networks have introduced a new business model for online e-commerce companies. Instead of selling news, products, or information, the typical social network sells information on its consumers to advertisers and other networks. In a sense, the users produce value by accidentally or on purpose, revealing their deepest desires, needs, wants, and status. The fact that a user lives in a certain place, belongs to a certain social group, has certain political leanings, and makes a certain amount of money is valuable information when collected, aggregated, and used to target individual users. Social networks automate extreme personalization and demographic targeting never before possible by any other media.

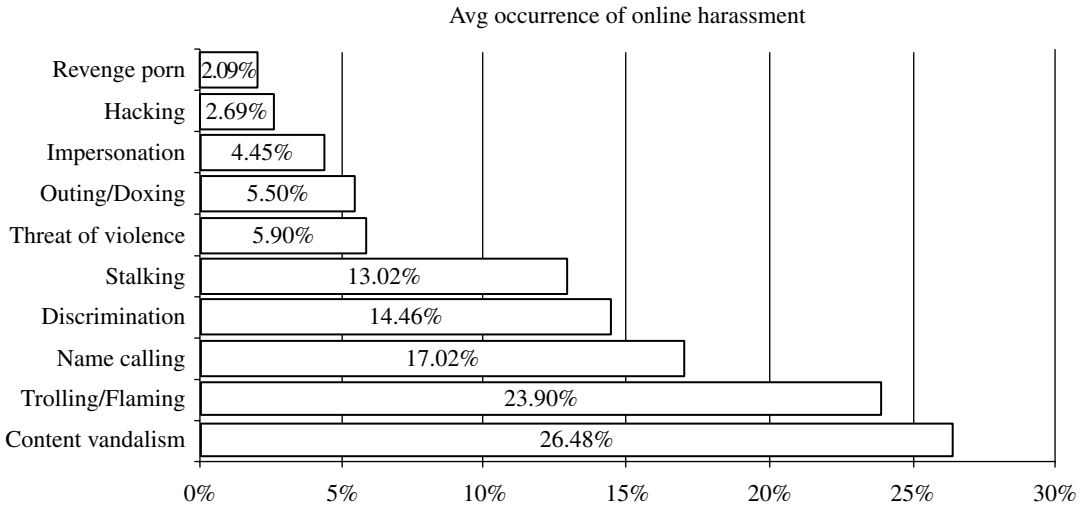


FIGURE 9.1 Forms of harassment experienced on Wikimedia. Source: Data from Ref. [10].

Social networks have inadvertently yielded an abundance of power to those with the ability to influence and persuade others. This power is mostly wielded for the good of the community. For example, Youtube.com is an excellent place to learn how to use new products, broadcast educational lectures from brick-and-mortar classrooms to billions of students throughout the world, and keep consumers up to date on breaking news. Collaborative network sites like Wikipedia.com pioneered a new way for humanity to self-organize and spread knowledge at light speed. Generally, an informed public makes for better citizenship and the social network has elevated the average person's ability to become informed.

But social networks have also yielded power to the dark side in the form of terrorist recruiting Web sites, negative social and political activism, harassment, and doxing—"the Internet-based practice of researching and broadcasting private or identifiable information (especially personally identifiable information) about an individual or organization. It is closely related to Internet vigilantism and hacktivism."[2]

The rise of online abuse parallels a corresponding rise of online social networks such as Twitter.com and Facebook.com. The Pew Research Center defines *online harassment* as any one of the following misbehaviors: being called offensive names, purposefully embarrassed, stalked, sexually harassed, physically threatened, and harassed in a sustained manner. Social activism promoted through media has a long history, but online social networks tend to magnify the impact, especially when harassment goes viral or false information transitions from an inflammatory post to fake news. The megaphone effect of a highly connected social network is something new in media and, as illustrated in this chapter, has been weaponized in the form of misinformation campaigns waged against individuals and nation-states.

In 2016 the Wikimedia Foundation Support and Safety team conducted a quantitative survey of Wikipedia harassment to determine types of abuse and its frequency of occurrence (see Fig. 9.1). They found that 54% of those who had experienced online harassment expressed decreased participation in the project where they experienced harassment. Online hate speech and cyberbullying are also closely connected to suppressing the expression of others, physical violence, and suicide. The list of abuses continues to grow beyond those enumerated in Figure 9.1 and spilled over into political and social actions generally classified as *fake news*. The ability to defraud social network users has become a national security problem because it undermines support for democratic institutions and may have political consequences such as swaying elections.

In 2018 alone, social network abuses expanded exponentially and brought unwanted attention from the US Congress.[3] Facebook.com changed its news feed algorithm to prioritize posts from friends and decrease postings from advertisers. This damaged Facebook.com's partners who experienced an 80% drop in referral traffic from Facebook and highlighted the power of a business model that makes money from collecting, analyzing, and selling its customer's data.

Billionaire investor and philanthropist George Soros called Facebook.com and Google.com a menace to society and called for tighter governmental regulation around the "monopolistic behavior of the giant IT platform companies." Facebook.com Chief Operating Officer Sheryl Sandberg ordered an investigation of Soros following his remarks.

[2]https://en.wikipedia.org/wiki/Doxing

[3]https://www.buzzfeednews.com/article/ryanmac/literally-just-a-big-list-of-facebooks-2018-scandals

Similarly, the government of Sri Lanka blocked Facebook.com and WhatsApp.com for three days after Facebook.com ignored calls to control ethno-nationalist accounts spreading hate speech against Muslims.

In March of 2018, trust in Facebook.com was further damaged, triggering a long sequence of backlashes against the social network when it was revealed that a political data analytics firm, Cambridge Analytica, had misused millions of Facebook.com user's data. The Cambridge Analytica incident resulted in a US congressional investigation and further reveals. Additionally, internal memos revealed an apparent lack of concern on the part of executives for the platform's impact on genocide in Myanmar and spate of suicides related to the social network. The Indian government claimed that at least 16 lynchings related to WhatsApp rumors led to 29 deaths in the country.

In September 2018 Facebook.com disclosed that the company had been hacked. The identity, email addresses, phone numbers, genders, locations, birth dates, and search histories of 30 million users were compromised. In December, the company announced that it had exposed photos of up to 6.8 million users affecting 1500 apps from 876 developers. In addition, a *New York Times* article reported the company had data sharing agreements with Amazon.com, Spotify.com, Netflix.com, Yahoo.com, and Microsoft.[4] This allowed partner companies to harvest Facebook.com data without user's knowing.

Also in 2018, Twitter.com revised its estimate of Iranian and Russia-backed posts going back to 2009. A Russian troll called the *Internet Research Agency* (IRA) was discovered to be spreading misinformation through various social networks.[5] Twitter.com, for example, revealed posts by 3841 accounts affiliated with the IRA and 770 other accounts potentially originating in Iran. The posts include more than 10 million tweets and more than 2 million images, videos, and Periscope.com broadcasts dating back to 2009. "This swath of social media is made up of a black marketplace of fake accounts, which actors ranging from the relatively harmless, like Coachella, to the nefarious—like Russian propagandists—can rent out. These accounts sit dormant until they are hired, and then spring into action, falsely amplifying tweets and hashtags so that more people see them."[6]

In February of 2018, a US grand jury indicted 13 Russians and the IRA on charges of violating criminal laws with the intent to interfere in US elections [1]. Based in St. Petersburg, the Internet troll was accused of engaging in online influence operations on behalf of Russian business and political organizations by spreading fake news through fake accounts registered in social networks. The company employed 1000 bloggers and commentators who daily posted 100 comments each. The IRA bloggers spread fake reports on an Ebola outbreak in Atlanta, Georgia, and a fake chemical plant explosion and sowed discord about the safety of vaccines.

The struggle over privacy and consumer data is not restricted to misinformation and propaganda among nations. In 2019 companies like Apple Inc., Facebook.com, and Google.com engaged in "data wars" over the collection and processing of consumer's data. Apple blocked apps from Google and Facebook, because both companies were collecting consumer data in a manner that Apple considered inappropriate. Apple claimed the tech giants violated Apple's developer program rules.

Google deployed an app called Screenwise Meter that collected data on a person's phone activity in exchange for gift cards. "Facebook distributed a market research app that gave the social network access to people's phone and web activity, paying them as much as $20 a month. The data Facebook could view included web searches, location data and even private messages" [2].

Apple suspended the use of Google and Facebook apps by deactivating their certificate server. When an app is activated, it must request authentication from an Apple certificate server before it is allowed to run on Apple's iOS operating system. A certificate server issues a pair of encryption keys—one for encrypting messages and the other for decrypting messages. The decrypting key is used to sign the app, therefore authenticating it. The signing process says that the app is approved by Apple and is authentic. Otherwise, iOS will refuse to run it on Apple equipment.

The conflict with Apple suggests something bigger—an impending data war among big tech companies vying for personal information on consumers. In this case, Apple was acting on behalf of consumers, but in general, companies that depend on consumer's aggregated data that is sold back to advertisers is a valuable asset that big tech companies may be willing to go to war over. It could be the beginning of corporate conflict that puts consumers in the middle.

Social networks are special because they evoke consumers to reveal more about themselves than other forms of online activity. An online bank only knows a consumer's name, address, telephone number, email address, and password. A social network knows an order of magnitude more. Typical social network databases contain abundant information about a consumer's likes and dislikes, friends, job, financial data, health data, locations, and so on. This rich trove of personal data is a double-edged sword. It can be used to facilitate searches and reduce the friction of online buying. It can also be used to target users for political purposes, spread of propaganda, and gaining access to bank accounts.

[4]https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html
[5]https://www.cnbc.com/2018/10/17/twitter-found-10-million-posts-by-iran-russia-backed-accounts.html
[6]https://qz.com/1429892/twitters-data-on-russian-bots-shows-how-much-of-the-platform-is-junk/

9.2 SOCIAL NETWORKS AMPLIFY MEMES

The power of social networks to propagate both true and false information without distinguishing one from the other is an inherent property of online communities principally due to constellations or tribal structures buried within the network. Figure 9.2 illustrates the tribe-like structure of a Facebook.com community linked together by *friending*. Tightly coupled structures form circular fan-like constellations that tend to magnify posts through spreading from user to adjacent user. This nearest-neighbor connectivity is epidemic-like and spreads memes at a rate determined by the number of nearest neighbors. That is, users with highly connected neighbors are more powerful than users with less connected neighbors. Spreading is accelerated by denser neighbors-of-neighbors interconnectedness.

Online social networks do not distinguish between types of memes—they spread both truth and falsehood with equal velocity and intensity. Misinformation and military-like misinformation campaigns pose a threat to national security and indirectly to the communication infrastructure sectors supporting governments at a political level and emergency management and law enforcement at operational levels. False reporting of E911 emergencies and misinformation regarding police activity are common examples of abuse of social networks. Influencing political activism and elections through false memes is another example of the power of highly connected users within a social network.

Denial-of-service malware evolved from digital viruses that spread and contaminate infected machines called *zombies* to sophisticated automatons called *bots*—short for robotic malware. A botherder controls bots from a distance, much like remote controlling a toy racecar. But bots evolved into more sophisticated malware from targeted infrastructure algorithms as found in Stuxnet to simulation of human users registered as real people online. These fake humans have their own photos and friends online and are difficult to distinguish from real people. Spread of misinformation through botnets, for political purposes, is a form of *computational propaganda*.

The further evolution of DDOS bots and botnets as they were married to social networks acerbates the problem of computational propaganda operating from within highly connected social networks. Sophisticated bots act like humans, look like humans, and outnumber humans in some instances. And they are tireless. The automation of influence over humans by bots embedded within social networks like Twitter.com became the subject of intense scrutiny following

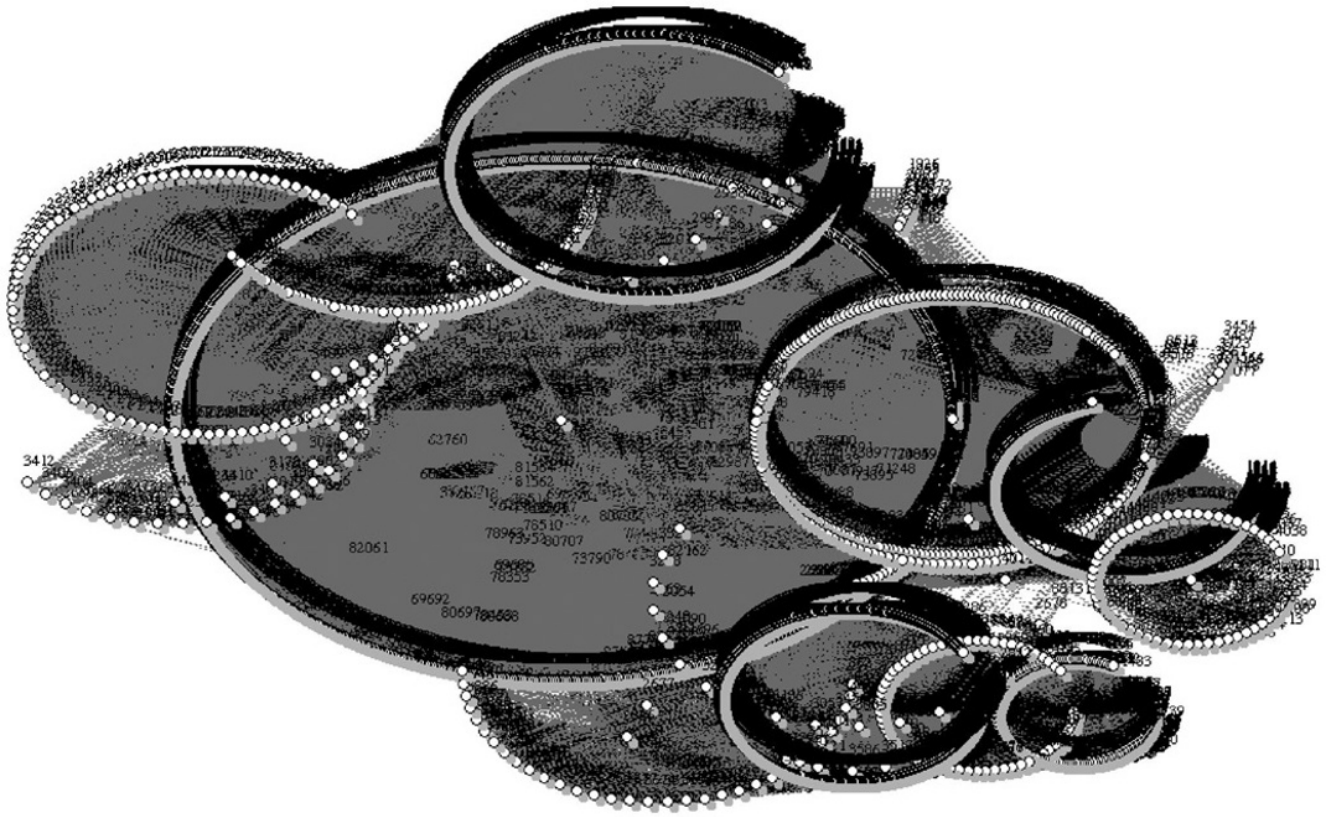


FIGURE 9.2 The nonuniform structure of a piece of the Facebook.com social network shows the formation of constellations or communities exhibiting tribal behavior. Source: Data provided by http://snap.stanford.edu.

the 2016 presidential election in the United States. Using proven techniques from the advertising world, botnets embedded within social networks became an effective propaganda weapon pulling human users into uncharted territory. Never before have psychological tools been combined with technical tools and shown to be massively effective across entire nations and even the entire globe.

A scientific study done by researchers at MIT found that fake news travels faster and further than truth. It also found that humans propagated emotion-laden fake news more than bots. According to a summary by the researchers, "We investigated the differential diffusion of all of the verified true and false news stories distributed on Twitter from 2006 to 2017. The data comprise ~126,000 stories tweeted by ~3 million people more than 4.5 million times. We classified news as true or false using information from six independent fact-checking organizations that exhibited 95 to 98% agreement on the classifications. Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information. We found that false news was more novel than true news, which suggests that people were more likely to share novel information. Whereas false stories inspired fear, disgust, and surprise in replies, true stories inspired anticipation, sadness, joy, and trust. Contrary to conventional wisdom, robots accelerated the spread of true and false news at the same rate, implying that false news spreads more than the truth because humans, not robots, are more likely to spread it" [3].

This research suggests that one way to separate fake news from truth is to observe how fast and far a meme travels over a period of time and quarantine it much like quarantining a contagious disease. In fact, the topological structure of the social network can be used to dampen and smother fake news.

9.3 TOPOLOGY MATTERS

The structure of a social network defines its topology. Some actors are more connected than others. Other nodes hold the network together, because without these blocking nodes, the network separates into islands. Still other nodes link with friends that have many friends—the friends of friends structure. Furthermore, topology determines effectiveness of computational propaganda—the use of misinformation to influence users. Figure 9.3 illustrates how topology affects four measures of influence spreading in a social network.

From Figure 9.3 it is easy to see how topology impacts virality of memes in social networks. Highly connected nodes, high betweenness centrality nodes and links, and high-influence nodes spread memes faster and further than lower-valued nodes and links. In simple terms, connectivity and centrality either facilitate or retard meme spreading. Centrality is increased due to blocking or clustering as shown in Figure 9.3b and c. Figure 9.3d confirms that topology matters by showing a heat map of meme spreading due to topological structure. The heat map was obtained by assuming a viral meme is passed on from one node to adjacent nodes with probability of 50%.

Bots and authentic human users within the social network take on network topological properties of connectivity, betweenness, and influence by nature of their position within the social network. Thus, the power of a tweet or "Like" depends on the actor's position within the network. Actors with high betweenness centrality and high connectivity have greater influence over other actors. When two posts contradict one another, the post with greater influence survives. The post with lower influence dies out, regardless of level of truth or falsehood. This dynamic is what dictates how far and fast memes travel. It is the underlying fundamentals of computational propaganda, echo chambers, and filter bubbles (described in Section 9.2).

Influence can be countered by greater influence. That is, tweets and posts have an implicit influence metric determined by network topology and the actor's position within the network. When opposing tweets and posts come into contact, it is the actor with the greatest topological influence that wins—the more influential post spreads faster and farther and therefore prevails over its competition. Which meme prevails across a social network is determined by the larger total influence an actor has, where total influence is the sum of actor influences that agree with the initiating post.

Consider the Facebook.com network of Figure 9.2 containing 4,039 nodes and 16,384 links. What happens when the most influential actor posts a meme that contrarians disagree with? Can the meme be stopped? Only 1.1% or an average of 45.4 contrarians are needed to stop an information cascade initiated by the most influential node (which is the secondary hub with 347 connections). As the size of the network grows, the number of contrarians required to halt the spread of influence emanating from an actor declines as a percentage of nodes in the network. Conversely, fake news can be successfully propagated throughout the entire network by recruiting sympathetic actors whose total influence exceeds the total influence of contrarians.

9.4 COMPUTATIONAL PROPAGANDA

Computational propaganda campaigns notoriously infiltrated Twitter.com networks to sway a California law on vaccination requirements. Twitter bombs smeared candidates in a 2009 special election in Massachusetts when nine fake user accounts produced 929 tweets within 138 minutes, triggering an *information cascade* intended to lend a sense of credibility

and grassroots enthusiasm to the fake news. Re-tweeting caused the Google.com search engine to promote the fake news to the top of its results page, further lending credibility to the misinformation.

Scientific American reports, "Real-world political outcomes are beginning to demonstrate the reach and power of bot-driven Twitter campaigns, in which a core group of tweets spreads information rapidly by encouraging large numbers of re-tweets. Recent investigations have uncovered, for example, Russia-backed bots programmed to automatically tweet animosity-stoking messages in the U.S. gun control debate following last month's school shooting in Parkland, Fla. That followed a wave of bots in January demanding (via the #ReleaseTheMemo campaign) the public release of a controversial House of Representatives document accusing the FBI of political bias in its surveillance activities during President Donald Trump's 2016 campaign. Just weeks after tweets hashtagged #ReleaseTheMemo went viral, Trump released

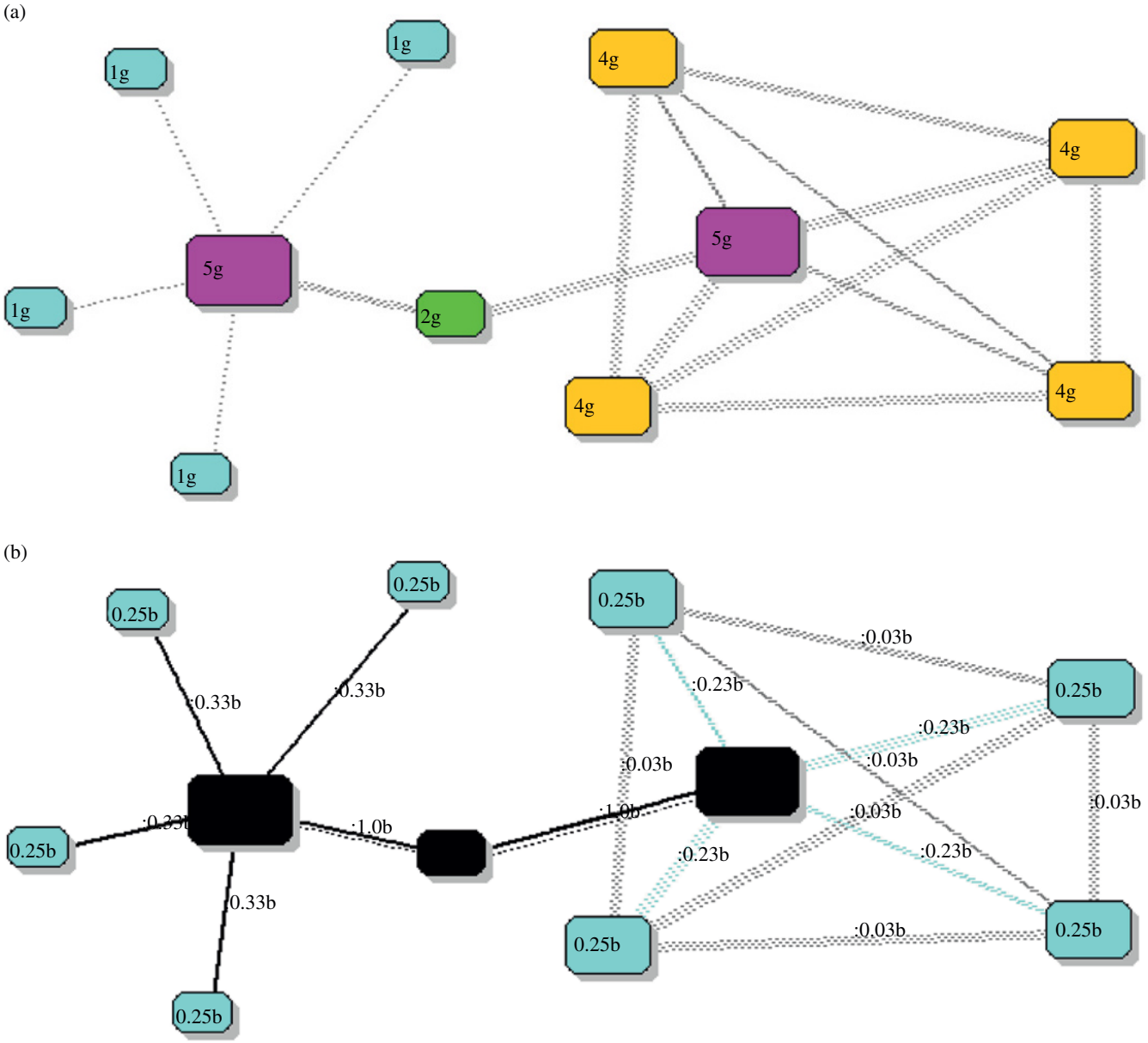


FIGURE 9.3 The topological structure of a social network determines how fast and far a social meme travels. (a) Connectivity is defined as the number of links connecting a node to other nodes. (b) Blocking nodes and links segment the network into islands if removed. Betweenness is a measure of influence a node or link has on the flow of information through the network. (c) An influence property called eigenvalue is an overall measure of influence each node has on other nodes. (d) The virality of nodes and links is a measure of how easily social memes travel from node to node through links.

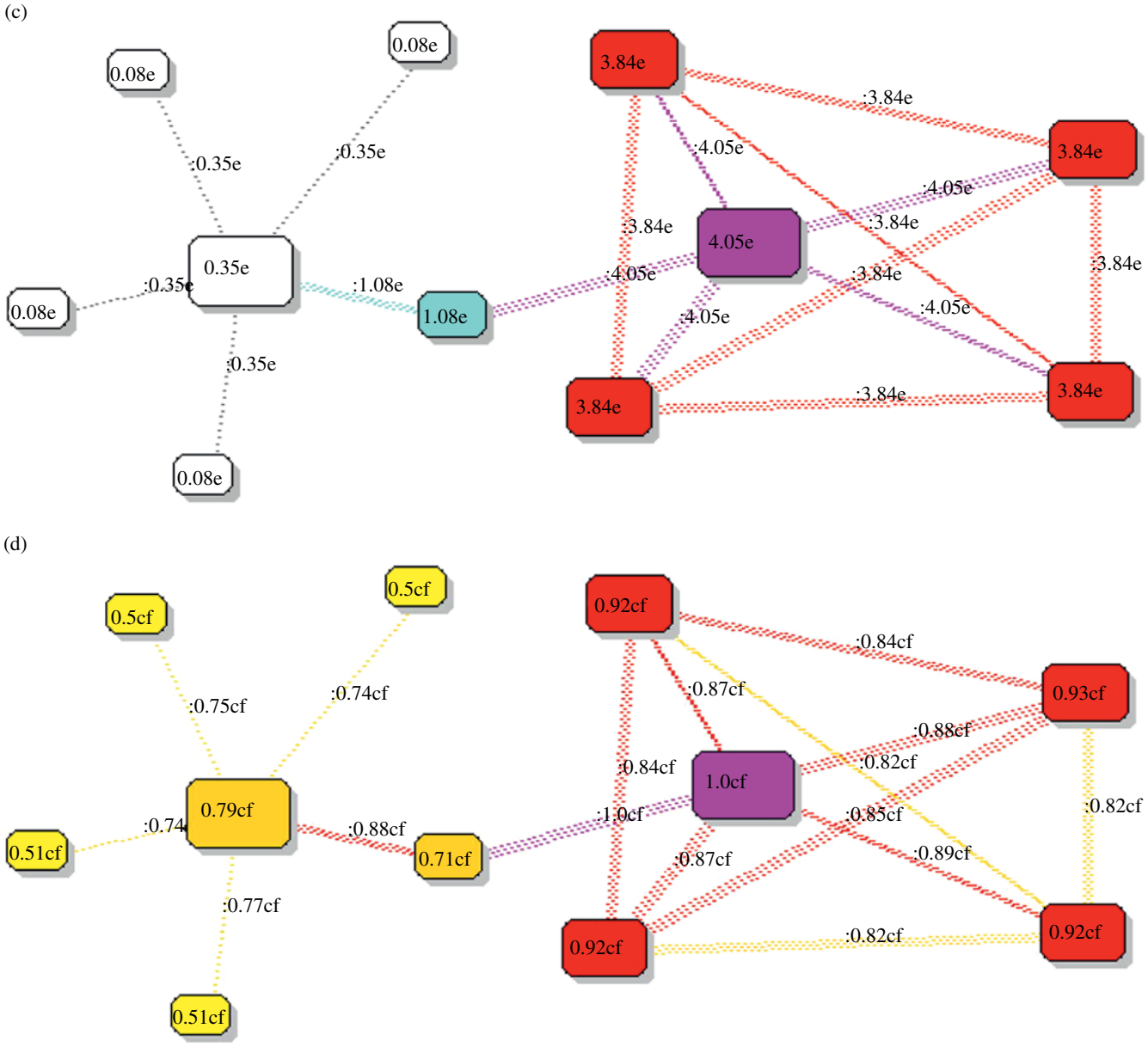


FIGURE 9.3 (Continued)

the memo—despite objections from the U.S. Department of Justice" [4].

A careful academic study of the 2016 US presidential election supports claims of election tampering: "The results of our quantitative analysis confirm that bots reached positions of measurable influence during the 2016 US election. Armies of bots allowed campaigns, candidates, and supporters to achieve two key things during the 2016 election: 1) to manufacture consensus and 2) to democratize online propaganda. Social media bots manufacture consensus by artificially amplifying traffic around a political candidate or issue. Armies of bots built to follow, re-tweet, or like a candidate's content make that candidate seem more legitimate, more widely supported, than they actually are. This theoretically has the effect of galvanizing political support where this might not previously have happened. To put it simply: the illusion of online support for a candidate can spur actual support through a bandwagon effect. … The goals of bot-driven tactics are manifold: to create a bandwagon effect, to build fake social media trends by automatically spreading hashtags, and even to suppress the opinions of the opposition. Bots allow for the democratization of digital propaganda because they make it possible for one person or group to massively enhance their presence online. Open APIs, and laissez-faire approaches to automation on sites such as Twitter, allow regular people to deploy their opinions en

masse. As one bot builder stated: if one person operating one profile can automate their profile to tweet every minute, just think what one person running one thousand automated profiles can do" [5].

Both political parties employed botnets to sway voter opinions. Clinton supporters deployed 13 hashtags and Trump supporters deployed 16 hashtags to spread their messages. They found 944 bots in the Trump botnet and 264 bots in the Clinton botnet. Betweenness centrality (bottleneck analysis) indicates "that bots also reached positions where they were able to control the flow of information between users. [Bots were] re-tweeted by humans, adding further evidence to the finding that bots influenced meaningful political discussion over Twitter, where pro-Trump bots garnered the most attention and influence among human users. Lastly, we provide preliminary evidence that bots were more actively involved in influencing the uptake of Trump-related hashtags than Clinton-related hashtags, with the potential to augment the megaphone effect."

Automated tools exist to determine the political orientation of Twitter users by examining the partisan preferences of their friends. Advertisers know that repetition and word-of-mouth communication is most effective. Built-in advertising tools let the misinformation campaigners exploit confirmation bias by amplifying what people already believe. Sending information multiple times from multiple sources increases the likelihood of acceptance. Fake memes become more believable with more repetition.

Gregory Berns, a distinguished professor of neuroeconomics at Emory University, illustrated the impact of groupthink on rational human beings in an experiment where subjects were asked to give an answer at odds with answers given by others.[7] For example, a photo of a dog is shown to a room full of people. Everyone asked to identify the animal says the dog is a cat. On average, uninitiated subjects went along with the incorrect group answer 40% of the time. Social pressure often causes people to change their picture of reality, and those that resist are emotionally upset.

Political astroturfing is the process of a single person or organization disguising a meme held by the person or organization as so-called grassroots activism. It is intended to make people think that everyone else believes a rumor or fake news, when in fact it is untrue [6]. An example of astroturfing through widely accepted advertising techniques is the Russian ad campaign in social media aimed at swaying the 2016 US presidential elections. Researchers noted that 470 ads were shared 340 million times and read by 10–20 million Facebook.com users during the campaign.[8] The Russians ads exploited Facebook.com communities labeled

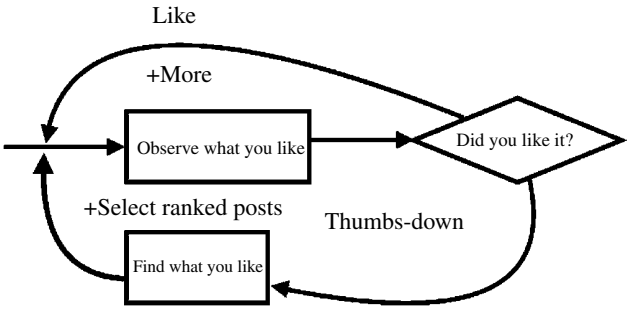


FIGURE 9.4 Filter bubbles are managed by software that rewards users with content they "Like" and selects content the software determines a user will like from psychological analysis of user behaviors. The diagram illustrates a kind of compulsion feedback loop employed by social network Web sites.

Blacktivists, Being Patriotic, Heart of Texas, KGBT United, Muslims of America, and Secured Borders. Armies of bots allowed campaigns, candidates, and supporters to achieve two key things during the 2016 election: (1) to manufacture consensus and (2) to democratize online propaganda.

9.5 THE ECHO CHAMBER

An echo chamber is a metaphor for a situation in which information, ideas, or beliefs are amplified or reinforced by communication and repetition inside a communication system. It is an apt description of social networks that employ technology to purposely amplify and promote some memes and minimize others. The technology for amplification and repetition is well grounded in psychology and plastic brain research and often leads to the creation of *filter bubbles*—"a term coined by Internet activist Eli Pariser – [to describe] a state of intellectual isolation that allegedly can result from personalized searches when a website algorithm selectively guesses what information a user would like to see based on information about the user, such as location, past click-behavior and search history. As a result, users become separated from information that disagrees with their viewpoints, effectively isolating them in their own cultural or ideological bubbles. The choices made by these algorithms are not transparent. Prime examples include Google Personalized Search results and Facebook's personalized news-stream."[9]

Various psychological techniques combined with social network software for maximizing stickiness have been employed from the very early days of the World Wide Web. For example, the simple "Like" and thumbs-up buttons supply feedback to algorithms that filter what users see and hear to increase participation. Figure 9.4 illustrates the feedback mechanism used by most social networks circa 2010 and beyond. Software selects posts made by the most

[7]https://www.psychologytoday.com/blog/am-i-right/201404/the-astonishing-power-social-pressure

[8]https://www.bizjournals.com/sanjose/news/2017/10/05/facebook-russia-ads-propoganda-reach-data-report.html?ana=apple_jo_video

[9]https://en.wikipedia.org/wiki/Filter_bubble

connected friends and friends of friends for the user. The user rewards or ignores the posts via the "Like" button or some form of feedback that indicates pleasure or displeasure with the software-selected posts. The feedback loop rewards users with information they already agree with, which accelerates more desired feedback in an exponential increase in user satisfaction.

Former vice president of Facebook.com, Chamath Palihapitiya, explains the power of "Like" buttons and software for increasing online participation: "The short-term, dopamine-driven feedback loops that we have created are destroying how society works. No civil discourse, no cooperation, misinformation, mistruth. This is not about Russian ads. This is a global problem. It is eroding the core foundations of how people behave by and between each other."[10]

The first-order effect of filter bubbles is obvious—it reinforces shared memes and leads to further spreading of memes throughout the echo chamber formed by social network connectivity. But second-order effects are subtler. The brain plasticity hypothesis says that the human brain physically changes due to the intense satisfaction it derives from pleasurable repetition. The filter bubble rewires our brains.

Brain plasticity researchers consider the human brain a plug-and-play organ that appropriates sensory feedback in the form of Twitter.com feedback, emotional feedback via Facebook.com, and groupthink from friends and friends of friends. Online interaction affects humans as much or more than humans affect online interaction. Professor Gary W. Small of the University of California–Los Angeles says, "The current explosion of digital technology not only is changing the way we live and communicate, but is rapidly and profoundly altering our brains."[11] Nicholas Carr, author of *The Shallows: How the Internet Is Changing Our Brains*, said, "A lot of people will assume that if our brains can adapt, then our brains will adapt to the flow of information and all will be well. But what you have to understand about neuroplasticity is that the process of adaptation doesn't necessarily leave you a better thinker. It may leave you a more shallow thinker."[12]

9.6 BIG DATA ANALYTICS

Social networks collect billions upon billions of data on their users and friends of users. This information is aggregated and analyzed to feed into compulsion networks as described in Section 9.2. Uploading a photograph from a smartphone to Instagram.com, for example, also uploads the metadata attached to the photograph. Photo metadata typically includes the photographer's name, creation date and location, and any copyright information required to assert ownership. When combined with personal information held by a social network, people in the photograph can be identified, location tracking automated, and personal information deduced. Most other media also incorporate metadata in addition to the media itself. For example, recorded sound includes the location, time, and data of recording, in addition to who created the recording.

Metadata is swept up along with clicks, cursor rollover, and URLs visited (cookies) to create a demographic and psychographic profile of users. This aggregated information is valuable to advertisers, but it is also valuable to scammers and malicious hackers seeking to exploit users in spear phishing exploits, for example. Clearly, political campaigns buy metadata to influence targeted voters.

The sum total of data collected per user may exceed millions of characters, but when combined with millions of users, it becomes *big data*—extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. Big data is the raw material that goes into deep learning algorithms to classify and target users for multiple purposes. For example, consumers wealthy enough to afford an expensive product might be classified separately from middle-income consumers. Furthermore, the wealthy consumers may be divided into subcategories, depending on brand preferences or readiness to buy.

The job of big data analytics is to classify consumers. The means of classification is generally machine learning algorithms belonging to a class of artificial intelligence called deep learning through artificial neural network processing. While other algorithms are used daily, deep learning classification of consumers is the most provocative because of its ability to sort through extremely large data sets and separate the information into extremely thinly sliced categories. For example, deep learning classifiers are able to identify a person from her photo by classifying the entire photographic library of online networks such as Facebook.com and Instagram.com.

There exists a wide variety of deep learning methods, but the following description is limited to the special artificial neural network architecture based on CNNs first proposed by Yann LeCun in 1988. Amazon.com uses CNNfor generating product recommendations and Google.com uses CNN to find people by searching through users' photos. A CNN is considered convolutional because it processes extremely large data sets a chunk at a time, from start to finish, until the entire data set has been processed. Figure 9.5 illustrates a typical CNN with input/output neurons (sensors) and several hidden or interior layers of neurons. The definition of a neuron here is based on organic neurons found in human brains, but it is much simpler. Essentially, digital neurons compute the sum of weighted inputs and output a zero or one

[10]https://en.wikiquote.org/wiki/Chamath_Palihapitiya

[11]https://lifeboat.com/ex/bios.gary.w.small

[12]https://www.huffingtonpost.com/entry/internet-changing-brain-nicholas-carr_us_5614037de4b0368a1a613e96?ec_carp=8232756251150797622

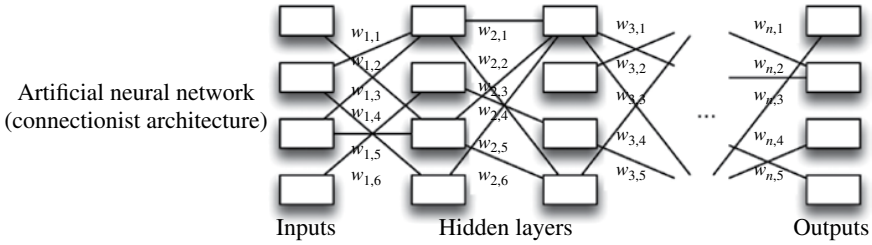


FIGURE 9.5 A typical CNN contains many layers of hidden neurons and robust interconnections between layers. Input signals are multiplied by weights as they flow left to right through layers to the output layer.

to the next neuron, depending on a threshold set by the CNN programmer. Thus, a neuron is a simple calculator for summing weighted input numbers.

The CNN is considered "deep" because it has many layers of neurons. Each layer is connected to the next layer through a network of weighted links. The weight's value is determined by a back propagation algorithm that minimizes the difference between the input and output values of a training set. Back propagation is an algorithm for matching a training set to known outputs so that the difference between inputs and outputs is minimized. The weighted sums at each neuron are compared with a threshold value. If the sum exceeds the threshold, the neuron emits a nonzero signal for the next layer to process and so forth until the output neurons are reached. Thus, the CNN is trained to recognize patterns that match the training set. Once trained, the CNN is able to classify new inputs according to the weights found for the training set.

The CNN may seem like a meaningless architecture, because it is not clear that sums converge to any meaningful values simply because a lot of data is processed. Intuitively, the larger the data set, the more likely the output patterns will be "blurred" or unrecognizable. But, in fact, the accuracy of classification improves with number of layers up to a point. (The best number of layers is unknown, in general, requiring a trial-and-error approach to CNN design.) As it turns out, the CNN must contain at least two hidden layers to separate input patterns into distinguishable output classes. (This is the so-called linear separability problem that was solved by adding layers.) In practice, a CNN may contain 15–30 layers.

Consider using a CNN to recognize photographs on Instagram.com. Suppose 10,000 photos are selected for training purposes. Each photo consists of pixels with a value determined by its color. The neurons in the input layer are assigned the pixel values, and the back propagation algorithm run to match output values with input values a closely as possible. This process is repeated for every one of the 10,000 training photos. Eventually, the weights assigned to the links between layers are established so that each layer matches a resolution-dependent outline of the each photo. The first layer may contain large blocks of color, the second layer may contain slightly more details, and the remaining layers contain gradually more enhanced details until the output layer looks like the input layer. Thus, an image is "recognized" by processing it into better resolution and greater detail layer by layer.

If enough layers are used and the neurons properly programmed, big data sets such as billions of photos from Instagram.com or billions of trades on the stock market are sliced and diced into categories. The input data is reduced to categories such that one set of inputs maps into a unique output class. The CNN is a clever machine for organizing large unstructured data sets into smaller and more organized classifications.

9.6.1 Algorithmic Bias

The matching mechanism of deep learning is problematic, however, and is a feature of CNN classification everyone should be concerned with. The main problem with CNN training is the likelihood of unanticipated results because of biases in the training data. This is known as *algorithmic bias* because it is introduced by the CNN, not the data. But, even without bias, CNN data harvesting is able to infer extraneous meaning from seemingly trivial user data. For example, big data analysis of dropped or fumbled smartphones says something personal and private about a person's risk when applying for a loan. Big data can make judgments about a person's suitability for a home loan, health insurance, and automobile insurance risk from seemingly unimportant data like how one mistreats her smartphone.

But what comes out of a CNN depends on what went into it. The CNN recognizes what it has been trained to recognize, and this is a more significant problem of big data analytics using CNN technology. Classification has been shown to introduce biases, such as misclassifying women as men, black people photos as animals, and convicted felons as repeat offenders. The algorithm may introduce sexual, racial, and gender bias itself. Thus, algorithmic bias has come to be directly associated with deep learning. The danger in big data analytics using CNN technology is that it might introduce "facts" that are not in the data. The CNN may generate fake data itself.

For example, a deep learning CNN is used by Google.com to automate the production of maps. Aerial photographs

are processed by a CNN called CycleGAN trained to recognize streets, buildings, and terrain.[13] It converts photos into street maps for navigation software. When inversely transformed back into photographs, Google employees realized that CycleGAN introduced data that was absent in the street map. Moreover, the added information often misrepresented the original information extracted from the photographs. Thus, deep learning has the ability to misinterpret its inputs even when trained under close supervision.

9.6.2 The Depths of Deep Learning

A second concern for deep learning is how adept it is at revealing second- and third-order details about a consumer's life. For example, most smartphones contain an accelerometer that measures jolts in the form of the pitch, roll, and yaw of the handset. Like an airplane flying through the air, pitch, roll, and yaw are measures of attitude along three degrees of freedom—nose up/down, body roll clockwise or counterclockwise, and headed left or right. As a user presses on the screen or keyboard of a smartphone, the accelerometer reports pitch, roll, and yaw. A deep learning CNN can process this navigational data and classify it according to which keys were pressed on the keyboard. A password or login can be determined by classification of accelerometer data.

> We show that accelerometer readings are a powerful side channel that can be used to extract entire sequences of entered text on a smart-phone touch screen keyboard. This possibility is a concern for two main reasons. First, unauthorized access to one's keystrokes is a serious invasion of privacy as consumers increasingly use smart phones for sensitive transactions. Second, unlike many other sensors found on smartphones, the accelerometer does not require special privileges to access on current smartphone OSes. We show that accelerometer measurements can be used to extract 6-character passwords in as few as 4.5 trials (median). [7]

9.6.3 Data Brokers

The collection and analysis of information scraped from social networks is a major industry mostly hidden in the shadows of the public square. Online consumers are unlikely to be aware of this infrastructure. In 2014, the Federal Trade Commission (FTC) of the United States described how this infrastructure works and warned of some of its potential pitfalls [8]. The following summarizes how brokers work:

- Data brokers collect consumer data from commercial, government, and other publicly available sources, largely without consumers' knowledge. These data include bankruptcy information, voting registration, consumer purchase data, Web browsing activities, warranty registrations, and other details of consumers' everyday interactions. Consumers are largely unaware that data brokers are collecting and using this information. Data brokers piece together data elements from multiple sources to form a detailed composite of the consumer's life.
- Data brokers provide data to each other: they provide data not only to end users but also to other data brokers. It is virtually impossible for a consumer to determine how a data broker obtained his or her data; the consumer would have to retrace the path of data through a series of data brokers.
- Data brokers collect and store a vast amount of data on almost every US household and commercial transaction. A typical database contains information on billions of consumer transactions, hundreds of billion of aggregated data elements, and trillions of dollars in consumer transactions. A typical database contains 3000 data segments for every consumer in the United States.
- Data brokers infer consumer interests from the data that they collect. They use those interests, along with other information, to place consumers in categories. Potentially sensitive categories include a focus on ethnicity and income levels, such as "Urban Scramble" and "Mobile Mixers," both of which include a high concentration of Latinos and African Americans with low incomes. Other potentially sensitive categories highlight a consumer's age such as "Rural Everlasting," which includes single men and women over the age of 66 with "low educational attainment and low net worths," while "Married Sophisticates" includes thirty-something couples in the "upper-middle class … with no children." Yet other potentially sensitive categories highlight certain health-related topics or conditions, such as "Expectant Parent," "Diabetes Interest," and "Cholesterol Focus."
- Data brokers rely on Web sites with registration features and *cookies* to find consumers online and target Internet advertisements to them based on their offline activities. Once a data broker locates a consumer online and places a cookie on the consumer's browser, the data broker's client can advertise to that consumer across the Internet for as long as the cookie stays on the consumer's browser. Consumers may not be aware that data brokers are providing companies with products to allow them to advertise to consumers online based on their offline activities. Some data brokers are using similar technology to serve targeted advertisements to consumers on mobile devices.

Data brokers depend on collection by others, most commonly through browsers and scraping of social network

[13] https://techcrunch.com/2018/12/31/this-clever-ai-hid-data-from-its-creators-to-cheat-at-its-appointed-task/

screens. Some browsers block pop-up ads, storing of cookies, and settings that prevent tracking of users as they click links and visit sites. The first line of defense, then, is to adjust settings on browsers to limit the amount of private information that is collected.

9.7 GDPR

The GDPR enacted by the EU became required by e-commerce sites operating in the EU in May 2018 following a long series of actions initially stimulated by the WikiLeaks exploit that stole information and computer security tools from the US National Security Agency (NSA) PRISM program. PRISM data collection is a highly controversial topic.

According to Wikipedia, "PRISM began in 2007 in the wake of the passage of the Protect America Act under the Bush Administration. The program is operated under the supervision of the U.S. Foreign Intelligence Surveillance Court (FISA Court, or FISC) pursuant to the Foreign Intelligence Surveillance Act (FISA). Its existence was leaked six years later by NSA contractor Edward Snowden, who warned that the extent of mass data collection was far greater than the public knew and included what he characterized as 'dangerous' and 'criminal' activities. The Guardian and The Washington Post published the disclosures on June 6, 2013. Documents indicate that PRISM is 'the number one source of raw intelligence used for NSA analytic reports', and it accounts for 91% of the NSA's Internet traffic acquired under FISA section 702 authorities. The leaked information came to light one day after the revelation that the FISA Court had been ordering a subsidiary of telecommunications company Verizon Communications to turn over to the NSA logs tracking all of its customers' telephone calls. U.S. government officials have disputed some aspects of the Guardian and Washington Post stories and have defended the program by asserting it cannot be used on domestic targets without a warrant, that it has helped to prevent acts of terrorism, and that it receives independent oversight from the federal government's executive, judicial and legislative branches. On June 19, 2013, U.S. President Barack Obama, during a visit to Germany, stated that the NSA's data gathering practices constitute 'a circumscribed, narrow system directed at us being able to protect our people.'"[14]

WikiLeaks triggered a series of actions largely spearheaded by Max Schrems, an Austrian lawyer who studied Internet privacy issues while attending the University of Santa Clara, California. Schrems advocated a strict set of regulations that eventually became known as GDPR. Subsequently, GDPR has spread across the Internet and has been enforced legally or voluntarily adopted by social network sites. The following details illustrate how social network activism is moving the Web closer toward regulation aimed at protecting consumer's privacy.

Section 230 of the Communications Decency Act of 1996 provides e-commerce companies with immunity from liability due to publication of information provided by users. Facebook.com and Twitter cannot be sued for something posted by users because they are considered an interactive computer service instead of a product company such as an automobile manufacturer. They are what have become known of as a "platform" unresponsible for what their users say and do. This became known as the *safe harbor* clause of the law and has been upheld by lower courts, but it does not protect users against federal criminal liability or intellectual property theft.

In 2013 Max Schrems filed a complaint against Facebook Ireland Ltd. with the Irish Data Protection Commissioner (DPC), Ireland being the country where Facebook has its European headquarters. The complaint was aimed at prohibiting Facebook from transferring data from Ireland to the United States, given the alleged involvement of Facebook USA in the PRISM mass surveillance program. Schrems based his complaint on EU data protection law, which does not allow data transfers to non-EU countries, unless a company can guarantee "adequate protection."

In October 2015, French magistrate Yves Bot serving as the presiding judge of the European Court of Justice declared the safe harbor agreement between the EU and the United States invalid on the grounds that the United States was not supplying an equally adequate level of protection against surveillance for data being transferred there. Subsequently, the European Commission and the United States agreed to establish a new framework for transatlantic data flows on February 2, 2016, known as the "EU–US Privacy Shield."

The EU–US Privacy Shield was a replacement for the International Safe Harbor Privacy Principles. US President Donald Trump signed an executive order entitled "Enhancing Public Safety," which states that agencies shall ensure that their privacy policies exclude persons who are not US citizens or lawful permanent residents from the protections of the Privacy Act regarding personally identifiable information. This did not satisfy the Europeans. German MEP Jan Philipp Albrecht and Max Schrems criticized the new ruling. Many Europeans demanded a mechanism for individual European citizens to lodge complaints over the use of their data, as well as a transparency scheme to assure that European citizens' data does not fall into the hands of US intelligence agencies.

The GDPR is the result of this dissatisfaction with handling of privacy information by social network e-commerce sites, especially Facebook.com. Broadly, GDPR guarantees a social network user's rights to be forgotten (deleted from the network), right to change one's mind and withdraw consent to use private information collected by the social network, child-friendly protections,

[14]https://en.wikipedia.org/wiki/PRISM_(surveillance_program)

and expedient security breach reporting (within 72 h). The GDPR defines personal information such as name, address, health records, biometric data, racial data, political data, and sexual preference.

At the time this was written, GDPR falls short of protecting user data obtained by sensors and aggregation of Internet of Things data. For example, GDPR does not prevent social networks from selling your anonymous psychographic profile to another e-commerce business, harvesting your cookies, recording your location, blocking fake news, or influencing users without their knowledge that they are subject to persuasion by unnamed advertisers. Harvesting cookies is especially egregious because cookies document URLs that have been visited; what a consumer looks at, bought, and recommended to friends; and GPS locations visited. Cookies contain enough personal information to build highly accurate profiles without GDPR-protected data.

9.8 SOCIAL NETWORK RESILIENCE

Figure 9.6 shows an adapted Hodges conceptual framework scoreboard for social network resilience. Recall the Hodges framework introduced in earlier chapters as a general framework for evaluating community fragility based on connectedness, stability, and sustainability. Hodges defines resilience in qualitative terms as the ability of a system to return to its original form or position or the ability to recover quickly from one or more exploits. In a social network, a fake news exploit, password attack, and politically motivated misinformation campaign by bots are all forms of exploits or attacks that challenge the resilience of the online community to function as intended. Such exploits threaten connectedness, stability, and sustainability of the social network community.

The adaptation of the Hodges framework to social networks assumes they are digital communities subject to fragility per the Hodges formulation. The framework is general and not specific to social networks. However, the author's proposed specialization of the general framework is appropriate for maintaining community resilience within a social network. Clearly, it is not the only specialization possible. The reader is invited to modify or extend Figure 9.6 and the following method of scoring the framework.

Each factor in Figure 9.6 is assigned a score from 0 to 100 indicating the level of resilience of social network relative to each factor. The sum of scores is normalized to fall between 0 and 1 by dividing by 900. The sum of dimension scores yields the overall score for the analysis. The scores for causal factors shown in Figure 9.6 were obtained from the author's analysis of the state of typical social network e-commerce sites such as Facebook.com, Twitter.com, and Instagram.com circa 2018, which proved to be a tumultuous year for social networks. Table 9.1 suggests that social networks circa 2018 were not very resilient in the face of causal factors identified by the author.

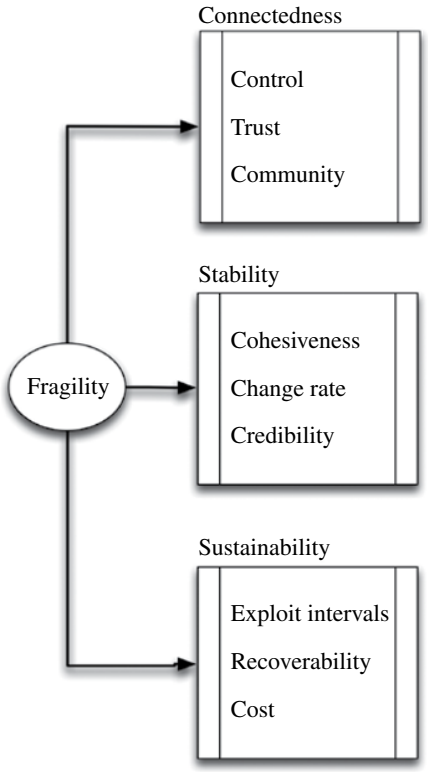


FIGURE 9.6 The Hodges conceptual framework adapted to social network resilience analysis consists of three dimensions (connectedness, stability, and sustainability), each with three causality factors.

Table 9.1 summarizes the author's scores applied to each causal factor that go into resilience. A score of 1.00 indicates maximum resilience, while a score of 0 indicates non-resilience. An overall total resilience score is obtained by simple summation. This produces an overall score between 0 and

TABLE 9.1 Resilience score for social network analysis using the Hodges conceptual framework

Dimension	Factor	Factor	Factor	Total
Connectedness	Control: 20	Trust: 20	Community: 90	130/900 = 0.14
Stability	Cohesiveness: 85	Change rate: 10	Credibility: 5	100/900 = 0.11
Sustainability	Exploit intervals: 0	Recoverability: 60	Cost: 75	135/900 = 0.15
Total				0.40

1.00 after normalization by dividing by 900. The overall score of 0.40 suggests a modest lack of resilience.

The rationale for each score is summarized here:

Control: 20. Social networks are facing loss of control over their business models, data collection practices, and independence from governmental intervention and regulation.

Trust: 20. Social networks are losing the trust of its community of users due to reports of loss of personal information, harvesting of big data, and selling consumer data to third parties.

Community: 90. Regardless of loss of control and trust, online communities have retained large audiences along with robust interaction among users.

Cohesiveness: 85. Communities have remained connected and functional even though users have experienced various forms of harassment, lost privacy, and been subject to the spread of fake news, propaganda, and other misinformation.

Change rate: 10. The rate of change in how a social network works (privacy settings, sharing of psychographic data on users) has been high.

Credibility: 5. The credibility of owners and operators of social networks was under attack by legislators during 2018 and likely continued. This raised concerns by consumers on the credibility of the e-commerce sites and their leaders.

Exploit intervals: 0. During 2018 the number of exploits expanded exponentially. Passwords were lost, political misinformation was spread, and attempts to distinguish between fake and factual news were largely a failure.

Recoverability: 60. Amazingly, most social networks recovered soon after each exploit was revealed. Monthly average users remained relatively constant even after the media reported massive failures and abuses.

Cost: 75. The cost to hacked social networks in terms of lost consumer data, loss of reputation, and technical updates was relatively modest as most social networks continued to operate with large margins. The stock price of most social networks varied little from overall stock market trends.

9.9 THE REGULATED WEB

A well-known security expert, Bruce Schneier says that we can no longer separate technology from everything else in our daily life. The tentacles are too embedded and pervasive. The question is, are we going to leave policy decisions up to corporations or governments? "Today, technology makes de facto policy that's far more influential than any law," he said. "Law is forever trying to catch up with technology. And it's no longer sustainable for technology and policy to be in different worlds" [9]. For instance, he said the Internet was never designed with any public policy in mind or with security in mind. Only researchers had access to it. Now it is critical to all aspects of our lives. It is more democratic, distributed, and commercial, and it moves a lot faster.

"Corporations have basically control over free speech and censorship regardless of laws," he said. "Corporations accept limitations on personal freedom because the technologies are our choice to use. They are for-profit systems. So now we hear terms like surveillance capitalism, algorithmic discrimination, digital divide, information attacks on democracies. These are not terms we heard even five years ago. And this means the Internet is no longer a separate thing. It's no longer its own world. It's part of consumer policy. It's part of automobile policy. … It's part of everything."

The implications are far reaching. Any technology that has pervasive influence on everything is likely to be abused. And abused technologies, such as nuclear power, gun control, and Internet, are destined to be regulated by government.

In the early years of e-commerce, Internet businesses were protected by safe harbor rules that separated the business from user-supplied content. A safe harbor social network was not held responsible for content supplied by its users. News and fake news were treated equally and supported by freedom of speech guarantees. But not all speech is free. Hate speech and threats of violence are generally prosecuted even in democratic societies. However, the distinction between freedom of expression and unacceptable expression may be narrowing. We may be entering the century of regulation.

9.9.1 The Century of Regulation

If the GDPR and algorithmic bias are leading indicators, the future of the Internet is regulation. What might such regulation be? The following is speculation based on the trend established by the GDPR and congressional hearings circa 2018. They are obviously subject to change. Generally, regulation centers on protecting data, limiting business models, and making advertising more transparent.

The prevailing wisdom of Web 2.0 social networks is that they are platforms that give voice to millions of people that previously had no voice. Global social networks like Facebook.com are good for democracy at large and individuals in the small. The social network enables people in a more powerful way than ever before. Reality is harsher: social networks have been used to damage democracy and given voice to terrorists as well as philanthropists. Social networks have been blamed for fomenting political discord in the Arab Spring uprising and for leading to the murder of innocent people in Myanmar. They have made life easier for consumers while robbing them of their privacy.

Like Prometheus, the Greek god that brought fire to humans, social networks have proven to be used for both good and evil. And like fire, humans must learn to tame social networks. The question is how to thread the needle between freedom of expression and censorship, between free enterprise and governmental control, and between individualism and groupthink. Left untamed, social networks are likely to tip toward nefarious uses more than beneficial uses.

Fortunately, there are analogs from nondigital sectors of civilized society that may show the way forward. First, freedom of expression is not unbounded. Most civilized societies place constraints on slander and hate speech. Hate speech is not regulated in the United States, because it is legally protected by the First Amendment. But it is regulated in Europe. As of January 2018, the Ministry of Justice of the EU enforces "NetzDG"—the Network Enforcement Act. The new law makes social media networks responsible for their users' content and fines tech companies up to €50 million (about $60 million) if they fail to remove "illegal" posts within 24 hours. The law impacts Facebook.com, Twitter.com, Google.com, YouTube.com, Snapchat.com, and Instagram.com but not networks like LinkedIn.com and its European counterpart Xing.com.[15] So, the bonds on free speech are being tightened, focusing on what is allowed on social networks.

Second, business models based on advertising have long been required to include transparency—what and who is paying for the ads and how consumer data may be aggregated and used. "Truth in advertising" should apply to online businesses just as it does to old media businesses. In fact, in the United States, the FTC requires truth in advertisement, "whether it's on the Internet, radio or television, or anywhere else, federal law says that ad must be truthful, not misleading, and, when appropriate, backed by scientific evidence. The Federal Trade Commission enforces these truth-in-advertising laws, and it applies the same standards no matter where an ad appears—in newspapers and magazines, online, in the mail, or on billboards or buses. The FTC looks especially closely at advertising claims that can affect consumers' health or their pocketbooks—claims about food, over-the-counter drugs, dietary supplements, alcohol, and tobacco and on conduct related to high-tech products and the Internet. The FTC also monitors and writes reports about ad industry practices regarding the marketing of alcohol and tobacco."[16] This regulation needs to be extended to fake news, use of botnets for political purposes, and the use of software-mediated filter bubbles.

Third, where does the responsibility lie for protecting consumer's privacy? In 2014 the FTC issued a report on data brokers that addresses many of the vulnerabilities described here. Recommendations forwarded to Congress are likely to be enacted into law as legislators become more aware of the issues. Specifically, regulation is likely to require social networks to:

- Give consumers access to their data.
- Allow consumers to opt out as a default.
- Control who has access to consumer data.
- Inform consumers on how their data is used.
- Require explicit permission to aggregate user data.

The report lists the most frequently collected data and who uses it:

- Online posts by consumers: data aggregators and brokers.
- Online shopping (clicks): e-commerce stores.
- Logins to register: Web sites and aggregators.
- Warranty and gift cards: e-commerce stores.
- Large purchases such as cars and houses: governments.

Finally, in 2019, California enacted a law requiring bots to declare themselves as bots and identify their botherders to social network consumers. "It is unlawful for any person to use a bot to communicate or interact with another person in California online with the intent to mislead the other person about its artificial identity for the purpose of knowingly deceiving the person about the content of the communication in order to incentivize a purchase or sale of goods or services in a commercial transaction or to influence a vote in an election."[17] The regulation applies to e-commerce and social network sites with 10 million or more unique monthly users.

9.10 EXERCISES

1. What does GDPR stand for?
   a. Global Data Privacy Rule
   b. Great Data Protection Regulation
   c. General Data Protection Rule
   d. General Data Protection Regulation
   e. Great Briton Privacy Regulation
2. What sets Web 2.0 apart from Web 1.0?
   a. 2.0 is two way.
   b. 1.0 was too slow.
   c. 1.0 was not secure.
   d. 2.0 uses IPv6.
   e. 2.0 started with social networks.
3. What is a social network meme?
   a. An unlawful idea that spreads
   b. An idea the spreads from person to person

[15]https://businessesgrow.com/2018/01/31/why-you-should-worry-about-europes-new-hate-speech-laws/

[16]https://www.ftc.gov/news-events/media-resources/truth-advertising

[17]https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

c. The selfish gene
d. The "me too" gene
e. A type of malware

4. Select the statement that is generally true of a social network.
   a. High centrality reduces the spread of memes.
   b. Low centrality increases the spread of memes.
   c. Low betweenness increases the spread of memes.
   d. High betweenness reduces the spread of memes.
   e. High connectivity increases the spread of memes.
5. Re-tweeting in Twitter can produce a (Select one):
   a. Twitter bomb
   b. Information cascade
   c. Spread of fake news to everyone
   d. Violation of privacy
   e. Meddling in elections
6. Political orientation of Twitter users can be determined by (Select one):
   a. Gregory Berns, professor of neuroeconomics
   b. Partisan preferences of friends
   c. Voting records
   d. Correlation with who won the election
   e. None of the above
7. What is political astroturfing?
   a. Twitter bombing
   b. Use of a disguised meme
   c. A Russian exploit
   d. An Iranian exploit
   e. A filter bubble
8. What is an echo chamber?
   a. A large room with poor acoustics
   b. A Web site for fake news
   c. What happens inside a filter bubble
   d. A psychological trick
   e. An example of the plastic brain hypothesis
9. The purpose of a CNN is to (Select one):
   a. Rewire our brains
   b. Reduce big data to simple facts
   c. Find and classify patterns
   d. Inject an algorithmic bias into memes
   e. Inject memes into a social network
10. Deep learning means (Select one):
    a. Finding deep meaning in big data
    b. Finding deep patterns in big data
    c. Reducing big data to rubble
    d. The ANN that has many layers
    e. The ANN that can recognize hidden patterns
11. True or false (Select all that are true)?
    a. A CNN may introduce extraneous meaning to simple data.
    b. Your password can be detected by reading accelerometer data from your phone.
    c. GDPR prevents the use of deep learning to discover passwords.
    d. GDPR outlaws cookies.
    e. Data brokers provide your data to each other.
12. Which of the following is enforced by GDPR (Select all that apply)?
    a. Your right to be forgotten.
    b. It bans cookies.
    c. It bans location tracking.
    d. It bans the spread of memes.
    e. It prevents foreign government meddling in US elections.
13. Which of the following is NOT part of the Hodges conceptual framework (Select all that apply)?
    a. Risk analysis
    b. Connectedness
    c. Stability
    d. Sustainability
    e. Cost-effective
14. The 1996 Communications Decency Act of the United States provides e-commerce with (Select one):
    a. Immunity from hacking
    b. Immunity from publication of information by users
    c. Immunity from the spread of memes
    d. Immunity from the GDPR
    e. An exemption from GDPR
15. Bots were deployed during the 2016 US presidential election by (Select all that apply):
    a. Democrats
    b. Republicans
    c. Communists
    d. People's Army
    e. Saudi Arabia

9.11 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. What can social networks do to prevent the spread of false memes? Include in your answer how you would protect freedom of expression and also protect consumers from fake news.

B. Deep learning and big data face a Promethean challenge of violation of privacy versus making consumer's lives easier and access to information and products more convenient. Propose a set of policies that accommodate both privacy and convenience.

C. Why is GPS location tracking via smartphones a concern? What harm might happen if a person's location is known to e-commerce sites like Google.com, Facebook.com, Instagram.com, and Twitter.com?

D. What is the (potential) harm in filter bubbles and the feedback of effect of "Like" buttons and other compulsion feedback loops employed by social networks?

E. Explain the plastic brain hypothesis and how human interaction with machines and social networks might alter the brains of large numbers of humans that spend a large proportion of their time online. What is the harm?

REFERENCES

[1] Mangan, D. and Calia, M. Special Counsel Mueller: Russians Conducted 'Information Warfare' Against US to Help Trump Win. *CNBC*, February 16, 2018.

[2] Wong, Q. and Shankland, S. *Apple's Clash with Facebook and Google: What You Need to Know*, February 2, 2019. Available at https://www.cnet.com/news/apples-clash-with-facebook-and-google-what-you-need-to-know/#ftag=CAD-09-10aai5b. Accessed July 23, 2019.

[3] Vosoughi, S., Roy, D., and Aral, S. The Spread of true and false news online, *Science*, 359, 6380, pp. 1146–1151. Available at http://science.sciencemag.org/content/359/6380/1146. Accessed July 23, 2019.

[4] Baraniuk, C. How Twitter Bots Help Fuel Political Feuds, *Scientific American*, March 27, 2018. Available at https://www.scientificamerican.com/article/how-twitter-bots-help-fuel-political-feuds. Accessed July 23, 2019.

[5] Wooley, S. C. and Guilbeault, D. R. Computational Propaganda in the United States of America: Manufacturing Consensus Online. Available at http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-USA.pdf. Accessed July 23, 2019.

[6] Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., and Menczer, F. Detecting and tracking political abuse in social media. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011, pp. 297–304.

[7] Owusu, E., Han, J., Das, S., Perig, A., and Zhang, J. ACCessory: Password Inference Using Accelerometers on Smartphones. Available at https://netsec.ethz.ch/publications/papers/owusu_ACCessory_hotmobile12.pdf. Accessed July 23, 2019.

[8] Ramirez, E. (Chairwoman), Brill, J., Wright, J. D., and McSweeny, T. (Commissioners). Data Brokers: A Call for Transparency and Accountability, May, 2014.

[9] Takahashi, D. and Schneier, B. It's Time for Technologists to Become Lawmakers. *VentureBeat*, March 6, 2019. Availableathttps://venturebeat.com/2019/03/06/bruce-schneier-its-time-for-technologists-to-become-lawmakers. Accessed July 23, 2019.

[10] Wulczyn, E., Nithum, T., and Lucas, D. Ex Machina: Personal Attacks Seen at Scale, *International World Wide Web Conference Committee (IW3C2), Published Under Creative Commons CC BY 4.0 License. WWW 2017*, April 3–7, 2017, Perth, Australia. ACM 978-1-4503-4913-0/17/04. http://dx.doi.org/10.1145/3038912.3052591.

10

SUPERVISORY CONTROL AND DATA ACQUISITION

SCADA is an acronym for *Supervisory Control and Data Acquisition*. SCADA systems are composed of computers, networks, and sensors used to control industrial processes by sensing and collecting data from the running process, analyzing that data to determine how best to control it, and then sending signals back through a network to adjust or optimize the process. A number of definitions exist. We chose the following, because it is operational and descriptive:

> An industrial measurement and control system consisting of a central host or master (usually called a master station, *master terminal unit* or MTU); one or more field data gathering and control units or remotes (usually called remote stations, *remote terminal units*, or RTU's); and a collection of standard and/or custom software used to monitor and control remotely located field data elements. Contemporary SCADA systems exhibit predominantly open-loop control characteristics and utilize predominantly long distance communications, although some elements of closed-loop control and/or short distance communications may also be present.[1]

The terms SCADA, ICS-SCADA (Industrial Control System-SCADA), EMS (Energy Management System), and DCS (Distributed Control Systems) are often used interchangeably, but the term SCADA is usually reserved for systems that are geographically dispersed. Because SCADA computers and networks monitor and control industrial systems, they are a special type of industrial control system (ICS). SCADA largely differs from more general enterprise systems because of its ICS mission—to provide automation services for industrial processes. In the following, SCADA and ICS-SCADA will be used interchangeably.

SCADA systems typically rely on communication networks to connect RTUs to the MTU. DCS, EMS, and Programmable Logic Controller (PLC) are various kinds of control systems that are similar to SCADA systems described here. However, according to industry experts, SCADA security is a discipline unto itself and requires special considerations. "SCADA is only one type of control system. The terms SCADA and DCS are not, and should not be used interchangeably."[2] For our purposes the distinction among the various kinds of control systems is not necessary because all kinds of systems are used in controlling critical infrastructures. Any control systems used in any critical infrastructure system that may render the sector vulnerable will be of interest in this book. In addition, many of the vulnerabilities of the Internet and enterprise IT systems are also vulnerabilities of SCADA systems because of digital convergence.

A thorough understanding of control systems is necessary because automation supports much of modern technological society. They run major portions of the transportation, energy, power, and water sectors as well as most manufacturing processes. If you have ever ridden in a subway, train, or automobile, your safety has been in the electronic hands

[1]http://www.sss-mag.com/glossary/page4.html

[2]A personal communication with Joe Weiss, April 2005.

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

of an SCADA system. Unfortunately, these systems are open to malicious software attacks much like the more general Internet and IT sectors.

This chapter discusses the following concepts:

- *SCADA, DCS, and other industrial control systems (ICS) are pervasive*: Automation of critical infrastructure sector processes found in water works, power, and transportation systems continues to increase, placing control in the hands of a machine—with human oversight. This trend will continue, making the study of control system security more relevant over time.
- *SCADA versus IT*: SCADA differs from enterprise IT systems largely because of their performance, availability, human safety, and centralization of assets, legacy drag, long lifecycles, and interdependencies. In general, ICS are more complex than consumer IT systems.
- *Responsibility is scattered*: NIST and NSA created the National Information Assurance Partnership (NIAP) to set standards and promulgate best practices, but there is no SCADA ISAC: rather industry is working with government within the Process Controls Security Requirements Forum (PCSRF). This chapter deals mainly with the NIST standards and recommendations, but a number of other recommendations are just as valid.
- *SCADA is vulnerable to cyber intrusion*: Because the components of SCADA systems are selected on the basis of low cost, their security has historically been sacrificed to reduce their cost and consumption of power. The result is that most SCADA systems are unprotected and so need to be hardened against cyber attack even more so than Internet devices such as personal computers, tablets, and phones. The lifecycle requirement of an SCADA system is typically 20–30 years, which means that upgrades are less likely to keep up with changing technology.
- *SCADA policies should focus on information assurance*: Because SCADA is generally vulnerable to asymmetric cyber intrusion, the focus of SCADA policy should be on cyber security and policies that reinforce best IT security practices. Safe practices are more likely to be effective than technical features.
- *Limits of redundancy*: Although duplication of equipment is expensive, SCADA systems (and IT systems in general) can be physically protected using redundancy of computers, communications, and facilities as illustrated by the case study in this chapter. However, redundancy can become a liability if redundant computers are connected to the same network because redundancy can spread worms and viruses farther and faster. SCADA systems are especially prone to the paradox of redundancy, because they often share the same computing base as the enterprise.

10.1 WHAT IS SCADA?

The purpose of all SCADA systems is to manage an industrial process—from monitoring a pipeline to adjusting the load in a power grid. Figure 10.1 shows a simplified representation of a typical SCADA system. These systems are operated through one or more Operation Control Centers (OCCs) containing computers, networks, and databases. The SCADA system database, for example, stores the state of the entire system—the condition of all sensors, valves, switches, and so on. It may contain the levels of all gas and oil storage tanks, pressures in water supply pipelines, or location and speeds of rapid transit systems under SCADA control.

Information about the state of the industrial system is obtained through a variety of sensors. Sensors may sense pressure, temperature, flow rates, and voltages. They may also be aligned with *actuators* to turn valves on or off, start and stop motors, and so on. The combination of sensor and actuator is called an RTU (*remote terminal unit*). The state of all RTUs is stored in the database and viewed through an OCC operator user interface—typically computer monitors, big-screen displays, and switches and dials mounted on a wall.

SCADA is a type of remote control system, meaning the valves, gates, switches, thermostats, and RTUs being controlled are many miles away from the OCC. RTUs located close to the devices being controlled report back to the OCC through a network. The RTUs can collect data—the Data Acquisition part of SCADA—and accept commands from the OCC to open/close a valve in a water pipe or report leaks in an oil or gas pipeline. Thus, the network is bidirectional.

An alarm management system also runs off the database. It constantly evaluates the state of the system by processing records in the database and monitoring the data streaming in from the network. If a certain reading is out of bounds or exceeds a threshold, an alarm is tripped, alerting the human operators. For example, in a power generation plant control system, an alarm may sound when temperatures exceed a certain threshold or sensors attached to the power grid detect a failure in a power line. In an oil pipeline system, sensors may collect data regarding leakage and report the location to an alarm management system so that repairs can be ordered. A transportation control system such as found in a subway or light rail train might report dangerous conditions to the alarm management system to prevent collisions.

Data is managed in a hierarchical fashion in most SCADA systems. The raw data collected by an RTU is aggregated at the RTU itself and then passed to a substation where it is

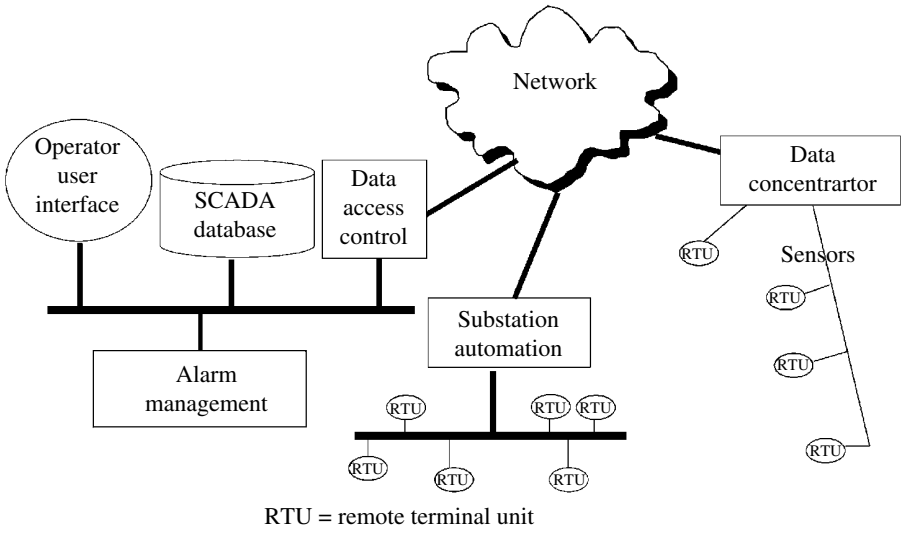


FIGURE 10.1 This diagram shows a simple view of a typical SCADA system and its components consisting of computers, networks, databases, RTUs, and software.

summarized or aggregated some more and then transmitted to one or more operation control centers where it is analyzed and summarized. SCADA collects data from geographically distributed sensors and delivers it to one or more processing servers. RTUs and OCCs can be distributed, especially if redundant computers are employed to increase reliability and security. This idea will be illustrated in a case study presented later in this chapter.

10.2 SCADA VERSUS ENTERPRISE COMPUTING DIFFERENCES

While SCADA shares many similarities with enterprise IT systems familiar to any user of the Internet, there are significant differences. The National Institute of Standards and Technology (NIST) special publication 800-82 enumerates the differences.[3] These differences are summarized here:

Performance requirements: SCADA requires real-time response, low latency, and high availability.

Availability requirements: ICS-SCADA systems must operate nonstop. Therefore, unanticipated outages are more consequential and maintenance outages must be planned and scheduled days/weeks in advance.

Risk factors: Business operation delays are important consequences as well as human safety. Major consequences are regulatory noncompliance, environmental damage, and loss of life, equipment, or production delays.

Architecture security focus: Focus is on protecting the IT assets and the information stored on or transmitted among these assets. Central server may require more protection because it is a critical hub.

Time-critical interaction: Response to human and other emergency interaction is critical. Access to ICS should be strictly controlled, but should not hamper or interfere with human–machine interaction.

Legacy drag: Legacy systems may not have desired features including encryption capabilities, error logging, and password protection.

Resource constraints: Systems are designed to support the intended industrial process and may not have enough memory and computing resources to support the addition of security capabilities.

Communication protocols: Communication protocols and media used by ICS communication are typically different from the enterprise IT environment and may be proprietary.

Patches: Software changes must be thoroughly tested and deployed incrementally throughout a system to ensure that the integrity of the control system is maintained. The ICS-SCADA system may use an operating system that is no longer supported.

Longer lifecycle: Component lifecycles typically on the order of 15–20 years. This makes it difficult to upgrade to new technology such as encryption and PKI.

Access to components: Components can be isolated, remote, and require extensive physical effort to gain access to them.

According to NIST special publication 800-82, "Initially, ICS had little resemblance to IT systems in that ICS were isolated systems running proprietary control protocols using specialized hardware and software. Widely available,

[3]http://csrc.nist.gov/publications/nistpubs/800-82/SP800-82-final.pdf

low-cost Internet Protocol (IP) devices are now replacing proprietary solutions, which increases the possibility of cyber security vulnerabilities and incidents. As ICS are adopting IT solutions to promote corporate connectivity and remote access capabilities, and are being designed and implemented using industry standard computers, operating systems (OS) and network protocols, they are starting to resemble IT systems. This integration supports new IT capabilities, but it provides significantly less isolation for ICS from the outside world than predecessor systems, creating a greater need to secure these systems. While security solutions have been designed to deal with these security issues in typical IT systems, special precautions must be taken when introducing these same solutions to ICS environments. In some cases, new security solutions are needed that are tailored to the ICS environment."[4]

Digital convergence is the most serious threat to SCADA systems going forward, because the TCP/IP monoculture increases the likelihood of cascade failures across multiple CIKR. Malicious software designed to attack Web sites will increasingly be able to attack ICS, banking, transportation, energy and power networks, and water supply systems through their SCADA networks.

10.3 COMMON THREATS

The hazards facing SCADA systems are different from those facing general enterprise systems. For instance, successful attacks on SCADA systems typically require more specialized knowledge of the industry served by SCADA. The following examples of intended and unintentional attacks on SCADA systems illustrate variety and ingenuity of the threat. These are summaries of detailed accounts reported by NIST:

Worcester air traffic.[5] In 1997, a teenager in Worcester, Massachusetts knocked out phone service at the control tower, airport security, airport fire department, weather service, the tower's main radio transmitter and transmitter that activates runway lights, and carriers that use the airport using the airport's unprotected dial-up modem.

Maroochy Shire sewage spill.[6] In 2000, a disgruntled rejected employee altered electronic data for sewerage pumping stations and caused malfunctions in their operations, ultimately releasing about 264,000 gallons of raw sewage into nearby rivers and parks using a radio transmitter to remotely break into the SCADA controls of the sewage treatment system.

Stuxnet worm.[7] Stuxnet is perhaps the most famous ICS hack aimed at disabling Iran's uranium processing centrifuges. Discovered in July 2010, the worm contains a highly specialized malware payload to target only specific SCADA systems—the control systems of centrifuges and control software made by Siemens Corporation.

CSX train signaling system.[8] In August 2003, the Sobig computer virus infected the computer system at CSX Corp.'s Jacksonville, Florida, headquarters, shutting down signaling, dispatching, and other systems. Trains between Pittsburgh and Florence, South Carolina, were halted because of dark signals, and one regional Amtrak train from Richmond, Virginia, to Washington and New York was delayed for more than 2 h.

Davis–Besse nuclear power plant.[9] In January 2003, the SQL Slammer worm affected the control networks of at least five power utilities including the private computer network at the idled Davis–Besse nuclear power plant in Oak Harbor, Ohio. It disabled the safety monitoring system for nearly 5 h and stalled the plant's process computer for 6 h.

Northeast power blackout.[10] One of the causes of the widespread power outage in August 2003 was traced to a failure of the alarm processor in FirstEnergy's SCADA system. It prevented control room operators from having adequate situational awareness of critical operational changes to the electrical grid serving 55 million consumers across Canada and the Northeastern United States. Additionally, effective reliability oversight was prevented when the state estimator at the Midwest Independent System Operator failed due to incomplete information on topology changes, preventing contingency analysis. At nearly the same time, several 345 kV transmission lines in Northern Ohio tripped due to contact with trees. This initiated cascading overloads of additional 345 and 138 kV lines, leading to an uncontrolled cascading failure of the grid. A total of 61,800 MW load was lost as 508 generating units at 265 power plants tripped.

The Zotob worm.[11] In August 2005, Zotob and its mutations caused computer outages at Caterpillar Inc., Boeing,

[4]ibid, section 3, pp 1.

[5]http://www.cnn.com/TECH/computing/9803/18/juvenile.hacker/index.html

[6]http://www.theregister.co.uk/2001/10/31/hacker_jailed_for_revenge_sewage/

[7]http://en.wikipedia.org/wiki/Stuxnet

[8]http://www.cbsnews.com/stories/2003/08/21/tech/main569418.shtml and http://www.informationweek.com/story/showArticle.jhtml?articleID=13100807

[9]http://www.securityfocus.com/news/6767

[10]http://www.oe.energy.gov/DocumentsandMedia/BlackoutFinal-Web.pdf

[11]http://www.eweek.com/article2/0,1895,1849914,00.asp and http://www.computerwire.com/industries/research/?pid=750E3094-C77B-4E85-AA27-2C1D26D919C7

and Daimler's US automobile manufacturing plants in Illinois, Indiana, Wisconsin, Ohio, Delaware, and Michigan.

Taum Sauk water storage dam failure.[12] In December 2005, the Taum Sauk water storage dam accidently released a billion gallons of water when the reservoir overflowed due to a pump failure. Apparently, the RTU gauges at the dam read differently than the gauges at the Osage plant located miles away at the Lake of the Ozarks.

Bellingham, Washington, pipeline failure.[13] In June 1999, 237,000 gallons of gasoline leaked from a 16″ pipeline and ignited 1.5 h later causing 3 deaths, 8 injuries, and extensive property damage. The pipeline failure was exacerbated by the SCADA system's poor performance that inhibited the human controllers from seeing and reacting to an abnormal pipeline operation.

Penetration testing incident.[14] A natural gas security consultant hired to perform security penetration testing accidently ventured into a part of the corporate network directly connected to the SCADA system. The penetration test locked up the SCADA system so that the utility was unable to send gas through its pipelines for 4 h.

What do these incidents have in common? First, they all occurred because of the complexity of ICS-SCADA systems. Second, they are all examples of normal accident theory (NAT) where a series of mistakes or perpetrated attacks combine and "snowball" into a much larger consequence. The 2003 power outage is a textbook perfect example of NAT, because one thing (SCADA fault) led to another thing (tripped lies), they magnified consequences as the rolling blackout spread across the Northeast. And finally, these ICS-SCADA incidents were unanticipated largely because of their complexity and hidden or obscured linkages.

10.4 WHO IS IN CHARGE?

Historically, PDD-63 did *not* specifically reference SCADA as a critical infrastructure, nor did it name SCADA as a component of other infrastructure sectors. Rather, cyber and physical security were given equal weight. According to PDD-63, "Critical infrastructures are those physical and cyber-based systems essential to the minimum operations of the economy and government." The Homeland Security Act of 2002 (H.R. 5005) assigned responsibility for information security to the Under Secretary for Information Analysis and Infrastructure Protection, which in turn created the National Cyber Security Division (NCSD) to address cyber security within critical infrastructure systems. Over the years, SCADA has held a special place within the IT sector due to its fundamental importance to almost all other sectors. SCADA is special, because it permeates manufacturing, transportation, healthcare, energy and power, and water and water treatment and continues to expand into other areas of modern civilization.

Section 225 of the Homeland Security Act of 2002 (CYBER SECURITY ENHANCEMENT ACT OF 2002) specifies penalties for cybercrime (up to 20 years for doing harm and life imprisonment for attacks that result in death), requires the DHS Under Secretary to report on cybercrimes to Congress, and outlaws Internet advertising of devices that may be used in cyber attacks.

The Homeland Security Act falls short on details for preventing attacks on computer systems—whether they are SCADA or standard information technology systems used by government, business, or consumers. The responsibility for establishing standards and guidance has been delegated to a combination of agencies and industrial groups. The NIST and the National Security Agency (NSA) have partnered to fill in the details concerning information security. This partnership is called the National Information Assurance Partnership (NIAP).

According to NIST, NIAP is a partnership between NIST and NSA:

> The National Information Assurance Partnership (NIAP) is a U.S. Government initiative designed to meet the security testing, evaluation, and assessment needs of both information technology (IT) producers and consumers. NIAP is collaboration between the National Institute of Standards and Technology (NIST) and the National Security Agency (NSA) in fulfilling their respective responsibilities under the Computer Security Act of 1987. The partnership, originated in 1997, combines the extensive security experience of both agencies to promote the development of technically sound security requirements for IT products and systems and appropriate metrics for evaluating those products and systems. The long-term goal of NIAP is to help increase the level of trust consumers have in their information systems and networks through the use of cost-effective security testing, evaluation, and assessment programs. NIAP continues to build important relationships with government agencies and industry in a variety of areas to help meet current and future IT security challenges affecting the nation's critical information infrastructure.[15]

NIAP has further delegated responsibility for working with the private sector to the Process Control Security Requirements

[12]http://en.wikipedia.org/wiki/Taum_Sauk_Dam_Failure

[13]www.ntsb.gov/publictn/2002/PAR0202.pdf

[14]http://www.sandia.gov/scada/documents/sand_2005_2846p.pdf

[15]http://niap.nist.gov/

Forum (PCSRF), an industry group organized under the National Information Assurance Program (NIAP). The members of this public–private organization are EPRI, American Gas Association (AGA), Association of Metropolitan Water Agencies (AMWA), and the Society of Instrumentation, Systems, and Automation (ISA). Government participation comes from NSA, DOE, and NIST. The so-called Common Criteria for Information Technology Security Evaluation, also known as the ISO/IEC 15408 standard, is being used to document the results of the NIAP effort.

The Common Criteria is not prescriptive. Instead, it is a process very similar to the risk assessment process described in Chapter 1. Common Criteria recommends certain documentation standards such as identification of threats, vulnerabilities, and risks associated with ICS and mitigations. At its roots, the ISO/IEC 15408 standard is an interpretation of a number of other standards (IEEE X509), practices, and procedures.

The NIAP/PCSRF initiative serves to set standards for control systems. The IT-ISAC provides linkages among the private sector companies who have a vested interest in making their network and computer products secure. But these governmental and industrial groups do not specifically address the processing needs of vertical sector components such as water treatment plants, power generation plants, and traffic control networks. These verticals are served by their own ISACs or various governmental agencies, but the vertical ISACs typically lack SCADA expertise. For example, the Department of Energy (DOE), which is responsible for power and energy infrastructure protection, has its own SCADA initiative. DOE makes its own recommendations apart from the IT-ISAC, NIAP, and the power industry.

From the foregoing, we can see that responsibility for SCADA and control system security is scattered across governmental agencies and commercial groups. Multiple initiatives by government agencies overlap and often duplicate one another, that is, DHS, NIST, NSA, and DOE all seem to play similar roles. Because SCADA cuts across various infrastructure sectors, the private industrial groups also overlap and support dual programs. For example, the various ISACs such as the IT-ISAC, WaterISAC, and Electric Power ISAC (EP-ISAC) perform similar functions when it comes to SCADA.

SCADA standardization efforts are spread across commercial and nonprofit organizations as well as governmental partnerships like NIAP. In addition, W3C, World Internet Society, and the IEEE promote their own information technology standards, which may or may not address control system security. This adds to the confusion on where to go for authoritative information. Who is in charge? Many private and public groups claim responsibility for SCADA. But like SCADA itself, the "command and control" of SCADA protection is spread far and wide. It is everywhere.

At the time this was written, the Control Systems Security Program (CSSP) within the Cyber Security and Communications (CS&C) division of Department of Homeland Security (DHS) and NIST encapsulated a number of risk-informed methods and practices in Cyber Security Evaluation Tool (CSET). This software tool helps organizations identify risks and implement secure protocols for protecting the SCADA and other cyber assets. CSET guides network administrators through a set of best practices and government standards that improve the security of IT and ICS-SCADA networks.

According to DHS, CSET incorporates standards from NIST, North American Electric Reliability Corporation (NERC), International Organization for Standardization (ISO), U.S. Department of Defense (DoD), and others. CSET produces a prioritized list of recommendations for improving the cyber security of enterprise IT and ICS-SCADA systems. It generates a detailed report on areas for potential improvement from answers to a list of questions.

The following sample questions are asked by CSET:

> Does the organization establish policies and procedures to define roles, responsibilities, behaviors, and practices for the implementation of an overall security program?
>
> Does the organization define a framework of management leadership accountability that establishes roles and responsibilities to approve cyber security policy, assign security roles, and coordinate the implementation of cyber security across the organization?
>
> How does the organization monitor physical access?
>
> How does the organization screen individuals?

10.5 SCADA EVERYWHERE

Control systems such as SCADA are used in almost every kind of industry. Application is not limited to critical infrastructure sector processes. SCADA can be found at work in amusement parks and noncritical factories. Widespread adoption was driven by efficiencies and economies—two drivers that are important in almost all industries. Automation reduces labor costs and increases reaction time. But SCADA in complex CIKR systems is fundamentally high risk because it has ignored security for decades. Most SCADA networks are as open as the telephone system and as vulnerable as a telephone line.

Below is a sampling of applications where SCADA reduces costs and increases reaction time, by automating various industrial processes:

- Food manufacturing
- Pharmaceuticals manufacturing
- Discrete parts manufacturing
- Environmental controls monitoring

Auto manufacturing
Railways/transit operations
Monitor and control mail sorting
Lock and gate security
Money production
Naval ship onboard monitoring
Power generation DCS
Transmission grid management
Power distribution DCS
Automatic metering
Oil refinery control
Oil pipeline management
Gas production
Gas pipeline management
Gas distribution
Gas supply management
Automatic metering
Clean water treatment
Wastewater treatment
Water supply management
Dams/aqueducts/spillways
Transportation control—subways, trams, and people movers at airports
Highway monitoring and control
Automation of bridge controls

The pervasiveness and extent of SCADA applications are staggering. For example, the flow of electric power through 672,000 circuit miles of overhead high-voltage transmission lines is governed by independent control systems that run unattended and parallel to the lines. Eighty percent of the nation's power is generated by 270 utilities. Each utility can generate up to 50,000 data collection points. In addition, the major DCS-controlled power generation plants are connected to over 3000 public and private electric utilities and rural cooperatives that make up the electric power grid. The market for power plant control was $1.5 billion and growing at about 6% per year in 2003.[16]

The United States currently uses 3250 billion-kilowatt hours of electricity, annually. A large part of this is generated by consuming 94 quadrillion BTUs of energy piped through 409,637 miles of interstate pipelines. Most of these are monitored and controlled by SCADA. Without safe and secure SCADA, these other CIKR sectors would not function.

For example, consider the Pacific Pipeline, which originates near the oil fields of Bakersfield and runs 132 miles to energy-hungry southern California (see Fig. 10.2). This plumbing transports 130,000 barrels/day of heavy crude oil from Bakersfield in the north to the Los Angeles refinery district located on the Pacific Coast. A parallel fiber optic network also runs the length of the pipeline so that SCADA computers can scan the entire length of the pipeline 4 times/s. The computers are looking for pipeline leaks that could lead to breaks and oil spills. This pipeline crosses several earthquake faults between Bakersfield and Los Angeles, so breaks are highly likely.

According to the Newton–Evans Research Company, 75% of the world's gas and oil pipelines of 25 km or more in length are monitored and controlled by SCADA systems.[17] Spending on these SCADA systems exceeds $200 million annually and is growing 30% per year. SCADA reduces the operational costs of gas and oil delivery by automating surveillance and emergency management. As we shall see, it also opens the door for asymmetric attacks on the power and energy delivery system.

10.6 SCADA RISK ANALYSIS

If SCADA networks and ICS were as simple as the schematic in Figure 10.1 suggests, the vulnerabilities would be limited—perhaps even inconsequential. But in reality, SCADA networks are intertwined with corporate networks, vendor connections, business partner connections, related Web sites, accounting and business process applications, and corporate databases. In practice, most SCADA systems live in a messy world of interdependent information systems (see Fig. 10.3). This business complexity introduces risk.

Human access to SCADA networks has steadily grown as productivity needs have increased, the number of business partners has grown, and the ease of networking has prompted public utilities, energy companies, and power operators to connect everything to everything else. Communication has improved efficiency and lowered cost, but it has also opened SCADA to network intrusion. It has added more vulnerability to the infrastructures it was designed to enhance.

To make matters worse, most devices in SCADA networks are low cost and low powered—optimized to be deployed by the tens of thousands. The RTUs are often inexpensive microcomputers with limited memory. They are not designed to support impenetrable security. For example, they usually do not support difficult to crack encryption or employ expensive firewall equipment that can block unauthorized access. Many RTUs are accessible over a simple dial-up telephone—an access method that can be used by anyone from anywhere in the world. When passwords are used, they are often the default password set by the manufacturer. It is simpler to use the default than change passwords on thousands of RTUs.

SCADA networks employ nearly every form of communication from Internet, Public Switched Telephone Network

[16]ARC Advisory Group, http://www.ARCweb.com

[17]Newton-Evans Research, Baltimore, MD, http://www.newton-evans.com

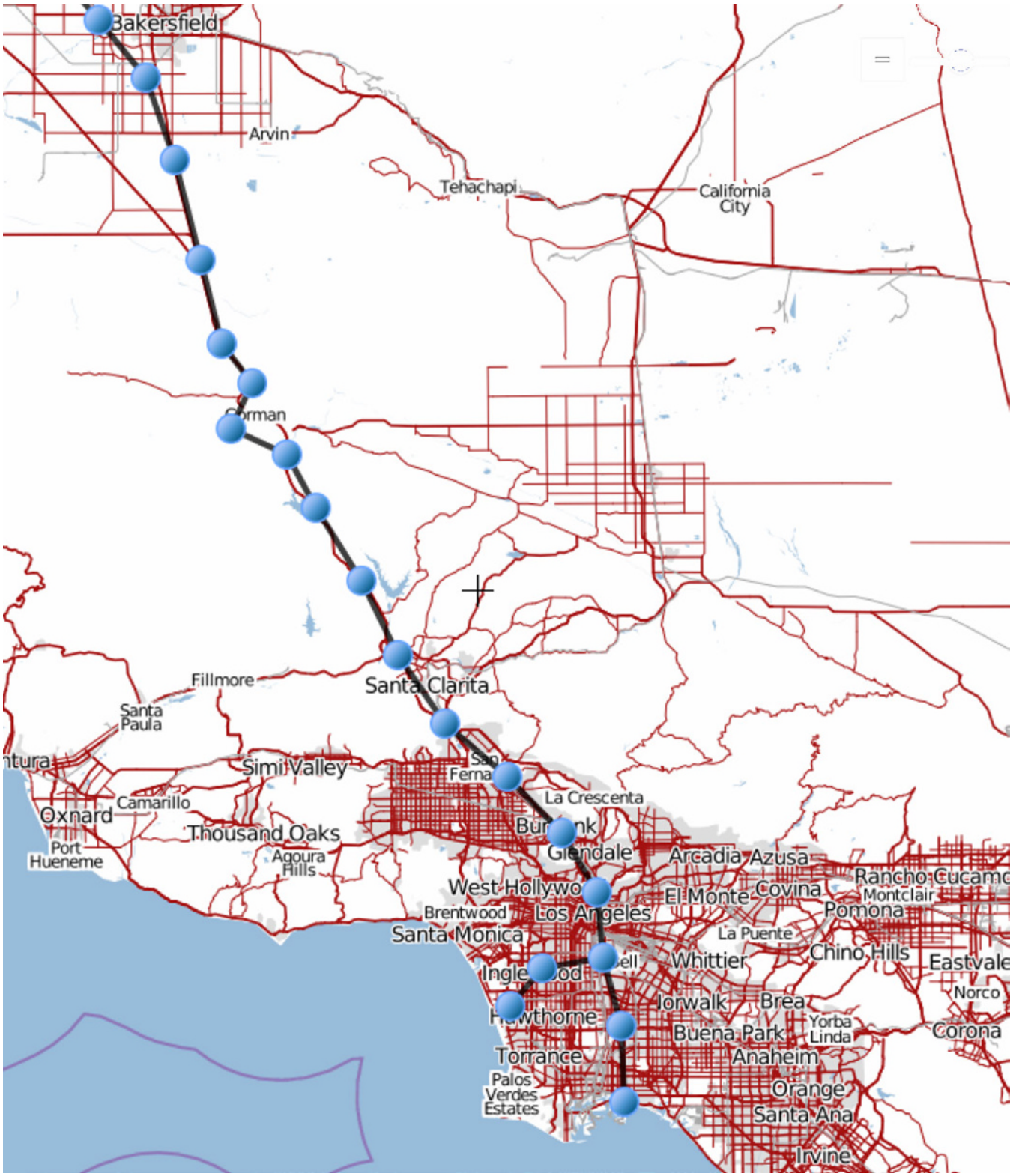


FIGURE 10.2 The 132-mile north-to-south Pacific Pipeline delivers crude oil from the oil fields of Bakersfield, California, to refineries on the coast next to Los Angeles.

(PSTN), Advanced Digital Network (ADN) (a form of PSTN similar to DSL), digital radio, digital satellite, and Wi-Fi wireless. All of these methods of communication have well-known security weaknesses. All are vulnerable to attack, and all are connected to the inner workings of the critical infrastructures they monitor.

The vulnerabilities of SCADA include the vulnerabilities of general information systems, plus additional SCADA-specific vulnerabilities:

1. Policy issues—Have best practices been put into place?
2. Business process problems—Are there vulnerabilities in the process itself?
3. System vulnerabilities—Are there vulnerabilities in the design of the system?
4. Open connectivity—Are there too many access routes that are unprotected?
5. Weak identification and authentication—Do users frequently change passwords?
6. Reliance on vulnerable technology—Is the technology itself vulnerable?
7. Protection outpaced by threat—Have your antivirus software and patches been updated?
8. Few security features built into technology—Is your equipment out of date?
9. In addition, SCADA and control systems generally must run 24 hours per day, every day, without failure.

Perhaps the biggest security hole in SCADA systems is traced to openness and connectivity with related internal business systems and external partners as illustrated in Figure 10.3. This openness has its advantages: business processes are made more efficient, and the resources needed to run SCADA systems can be shared with other IT functions. In addition, skills needed to maintain SCADA are not altogether different than general IT support skills.

The concentration of IT assets and streamlined networking of SCADA with other IT processes has its down side: it leaves SCADA vulnerable to denial of service attacks, Internet viruses, and malicious software. This is a familiar story—economic and competitive forces make it attractive for businesses to connect SCADA with everything else in the

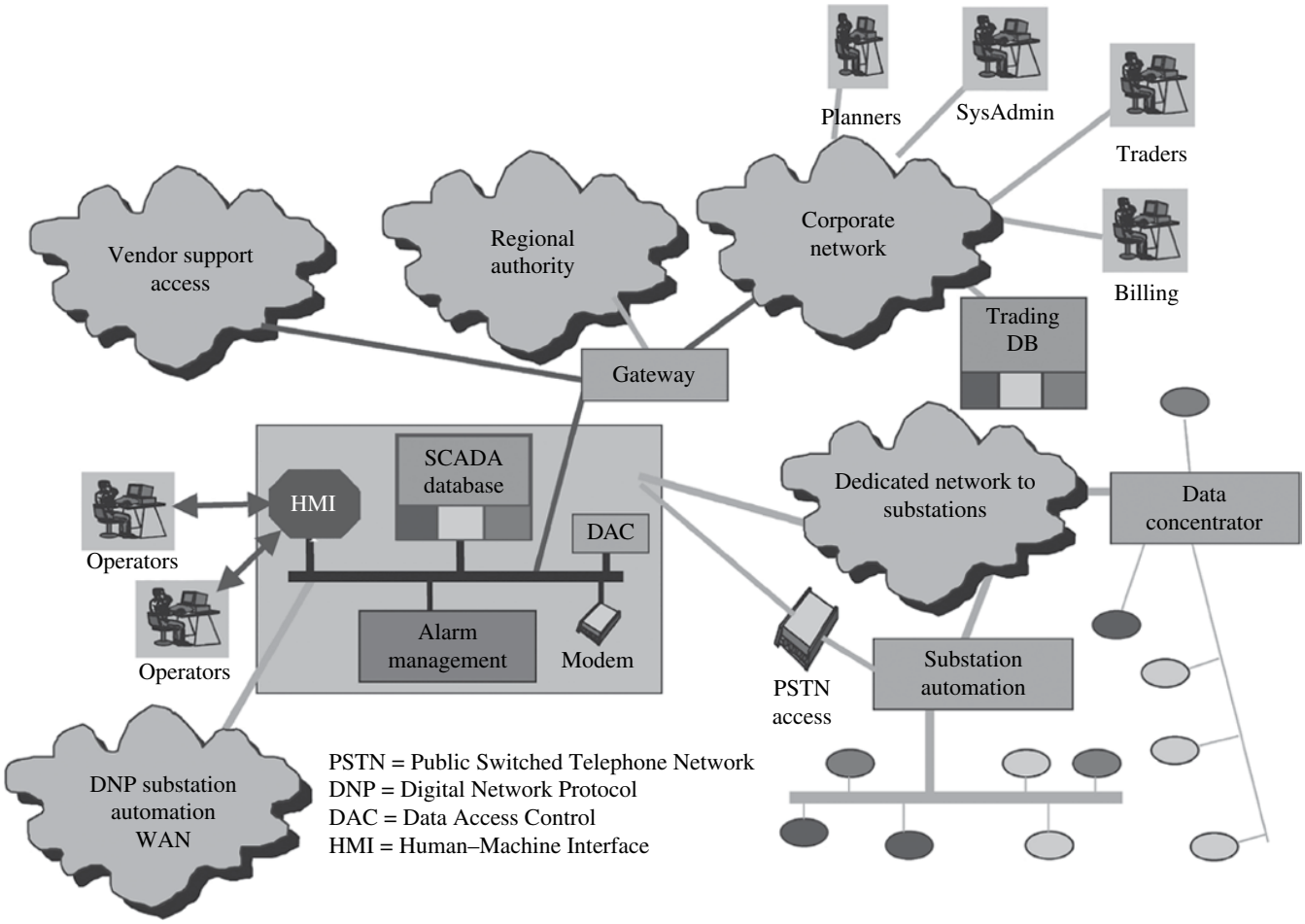


FIGURE 10.3 Most SCADA systems are open to access by a number of partners and vendors.

enterprise, and yet this is exactly the wrong thing to do if security is paramount. In addition, typical information technology workers may lack the specialized knowledge needed to secure SCADA components. SCADA is subject to self-organization just as other sectors are.

Figure 10.4 shows a general fault tree for SCADA risk analysis. The human error component of the fault tree lists major threats that lead to intrusion: operator error, hack attack, insider threat, and user error. The equipment failure component lists software flaws, component faults, physical attack, and power outage threats.

Only the detail under *human failure* is described here. The threats described under human failure are the most common weaknesses found in typical SCADA systems:

Remote access (REMOTE): Almost all SCADA systems allow dial-up connections via old-fashioned (but inexpensive) modems. In many cases, the dial-up connection is not protected and allows anyone to directly access an RTU or SCADA database or both.

System controls lacking or nonexistent (CONTROLS): The security of many SCADA systems have simply been overlooked or eliminated to save money. They are not protected against cyber attack because the necessary control software has not been implemented. Encryption, for example, introduces additional overhead and adds to the cost of a SCADA component.

Weaknesses in corporate networks (CORP-NET): Many SCADA systems are linked to corporate networks through a shared network link, a computer that is connected to both, or indirectly through dial-up communication lines. Thus, if the corporate network is penetrated, the SCADA network is indirectly penetrated.

No logging (LOGGING): System operators typically keep operational logs of events that have taken place during each work shift. The same logic applies to the SCADA system itself. Each data access—what kind of access was made and by whom—should be logged in a file. Unauthorized access should be denied, and the unauthorized attempt should be logged as well. Failure to keep logs and control access is like leaving the front door of your home open to burglars.

Vendors and partners (PARTNER): The so-called perimeter of an SCADA network is expanded because the access points have been extended by allowing more and more users to connect and access the databases maintained by SCADA. While this improves the efficiency of business processes, it also increases vulnerability. Business partners and vendors should be required to follow the same authentication and security procedures as employees.

Individual user authentication rarely enforced (AUTHENTICATE): Perhaps the most common vulnerability comes from the users themselves. Passwords are the first line of defense, and yet most users either do not use passwords, or they do not change them frequently enough to ward off password crackers. Passwords must be managed just like the keys to your car or house.

These (and other) threats are exploited through operator errors, hacker attacks, insider attacks, and user errors. Exact vulnerabilities (probability of successful attacks) for each of these generic threats are not generally known, nor are the financial damages resulting from a successful attack. This makes it difficult to estimate financial risk. Assuming maximum ignorance—threat and vulnerability are 50%—and a uniform consequence of $100 thousand, risk and overall vulnerability reduction follows the familiar exponential decline, according to Figure 10.4b. For example, risk is cut in half with an investment of $14 thousand.

10.7 NIST-CSF

DHS recommends a number of sector-specific risk assessment frameworks and methods. The NIST-CSF was introduced in Chapter 1. Recall that it is based on five steps: Identify, Protect, Detect, Respond, and Recover. Table 10.1 lists the items to check for the Protect step only.

While the NIST-CSF is not a risk-informed decision tool, it can assist in a risk assessment by assigning a numerical score to each checklist item. For example, each row of Table 10.1 might be assigned a number between 0 and 10, signifying how close each item comes to perfection. A risk index can them be calculated or the entire Protect step by subtracting the sum of scores divided by the total from one. For example, if scores sum to 150 out of a total of 500, the risk index is 1.0—150/500 or 0.70. This suggests the system is 70% vulnerable to one or more weaknesses.

10.8 SFPUC SCADA REDUNDANCY

Chapter 11 on water analyzes the San Francisco Public Utilities Commission (SFPUC) water supply system known as the *Hetch Hetchy*. The following is an analysis of the SFPUC water SCADA system that monitors and regulates the Bay Area's drinking water. This major metropolitan water and power system underwent a major upgrade in 1999–2000 to harden it against natural and manmade disasters. The analysis presented here uses hypothetical values to evaluate the resilience of the cyber component, while the next chapter evaluates the physical component.

In November 2002, San Francisco voters approved legislation to finance the largest renovation of a water delivery system in San Francisco history. The $3.6 billion capital program contained 77 projects to repair, replace, and seismically upgrade the water system's aging pipelines and tunnels, reservoirs, and dams. The first phase of the massive renovation amounted to $1.6 billion. In addition, a $10.5 million upgrade

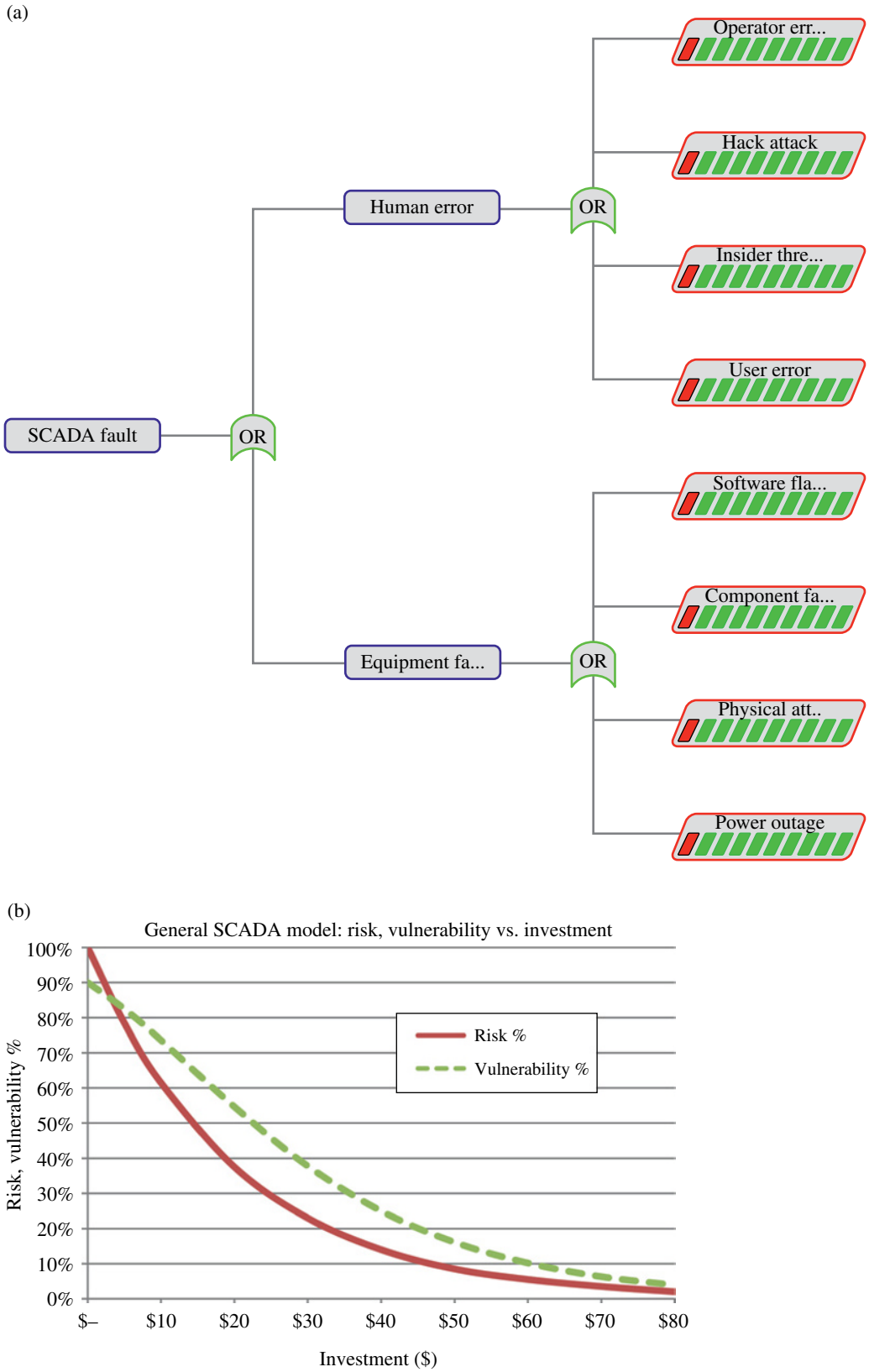


FIGURE 10.4 General fault tree of possible vulnerabilities of a typical SCADA system. (a) General SCADA fault tree showing typical threats from humans and system faults. (b) Risk and fault tree risk and vulnerability exponentially decline with investment.

TABLE 10.1 Checklist for the protection step of the NIST-CSF is composed of access control, awareness and training, data security, information protection, maintenance, and protective technology

AC: Access control
PR.AC-1: Identities and credentials are managed for authorized devices and users.
PR.AC-2: Physical access to assets is managed and protected
PR.AC-3: Remote access is managed
PR.AC-4: Access permissions are managed, incorporating the principles of least privilege and separation of duties
PR.AC-5: Network integrity is protected, incorporating network segregation where appropriate
AT: Awareness and Training
PR.AT-1: All users are informed and trained
PR.AT-2: Privileged users understand roles and responsibilities
PR.AT-3: Third-party stakeholders (e.g., suppliers, customers, partners) understand roles and responsibilities
PR.AT-4: Senior executives understand roles and responsibilities
PR.AT-5: Physical and information security personnel understand roles and responsibilities
DS: Data security
PR.DS-1: Data-at-rest is protected
PR.DS-2: Data-in-transit is protected
PR.DS-3: Assets are formally managed throughout removal, transfers, and disposition
PR.DS-4: Adequate capacity to ensure availability is maintained
PR.DS-5: Protections against data leaks are implemented
PR.DS-6: Integrity checking mechanisms are used to verify software, firmware, and information integrity
PR.DS-7: The development and testing environment(s) are separate from the production environment
IP: Information protection
PR.IP-1: A baseline configuration of information technology/industrial control systems is created and maintained
PR.IP-2: A System Development Life Cycle to manage systems is implemented
PR.IP-3: Configuration change control processes are in place
PR.IP-4: Backups of information are conducted, maintained, and tested periodically
PR.IP-5: Policy and regulations regarding the physical operating environment for organizational assets are met
PR.IP-6: Data is destroyed according to policy
PR.IP-7: Protection processes are continuously improved
PR.IP-8: Effectiveness of protection technologies is shared with appropriate parties
PR.IP-9: Response plans (Incident Response and Business Continuity) and recovery plans (Incident Recovery and Disaster Recovery) are in place and managed
PR.IP-10: Response and recovery plans are tested
PR.IP-11: Cybersecurity is included in human resources practices
PR.IP-12: A vulnerability management plan is developed and implemented
MA: Maintenance
PR.MA-1: Maintenance and repair of organizational assets is performed and logged in a timely manner, with approved and controlled tools
PR.MA-2: Remote maintenance of organizational assets is approved, logged, and performed in a manner that prevents unauthorized access
PT: Protective technology
PR.PT-1: Audit/log records are determined, documented, implemented, and reviewed in accordance with policy
PR.PT-2: Removable media is protected and its use restricted according to policy
PR.PT-3: Access to systems and assets is controlled, incorporating the principle of least functionality
PR.PT-4: Communications and control networks are protected

of the SFPUC water SCADA system was approved and implemented as described here. The purpose of this case study is to illustrate how SCADA systems like the SFPUC SCADA can be evaluated using the tools developed thus far in this book.

10.8.1 Redundancy as a Resiliency Mechanism

The SFPUC SCADA network reduced risk against single-point failures using *redundant components*. As this example illustrates, duplication of communication links, OCCs, and computer equipment goes a long way toward hardening SCADA against physical attacks or equipment faults. Redundancy may be one of the most effective methods of protecting other critical infrastructures, but IT sector redundancy may introduce additional risks due to the paradox of redundancy. Therefore, cyber security precautions must also be taken.

Redundancy dramatically reduces the probability of a fault because the probability of individual component failures is *multiplied* rather than added together. This is the nature of AND logic in the fault tree. Hence, if an individual component

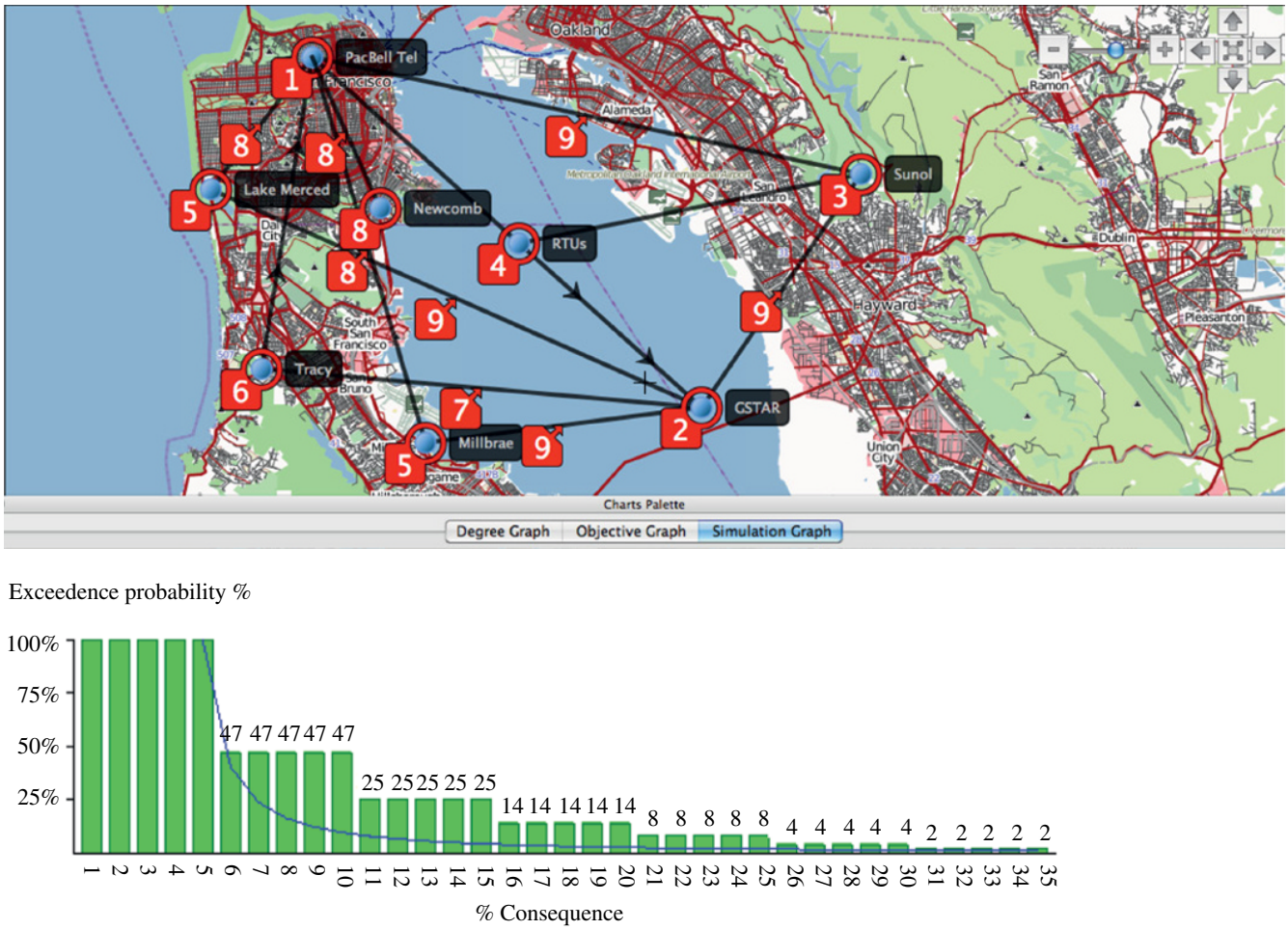


FIGURE 10.5 Major nodes of the SFPUC water SCADA network are connected by landlines, satellite communication, and one radio link. Nodes and links are ranked according to connectivity and betweenness.

fails with probability of 10%, two identical components fail with probability of 10% times 10 or 1%. Three identical components fail with probability of 10% times 10% times 10% or 0.1%. This dramatic reduction in vulnerability is more than a mathematical fact—it actually works in practice.

This is how the SFPUC SCADA system was hardened. Triple redundancy means each critical component of the SCADA system was duplicated three times. The resulting fault probability is decreased by three orders of magnitude. So instead of a 1% probability of failure, a triple redundant system has an extremely small 1-part-per-million failure probability (1%×1%×1% is 0.000001 or 1 part in 1 million).

The upgraded SFPUC water SCADA network is shown in Figure 10.5. The nodes of this network represent OCCs, RTUs (78 RTUs in the entire system are represented by one node in Fig. 10.5), and the major communication services—GSTAR satellite, digital radio, and traditional telephone lines.

Redundancy exists in the form of three OCC—one each at Tracey, Lake Merced, and Sunol. Data is distributed to all three OCCs, simultaneously. There are redundant servers and multiple workstations inside each OCC. The servers work from the same SCADA database, so there is always a backup server in case the primary server fails. Thus, failure in one OCC does not lead to overall failure, because operations are transferred to another OCC.

The communication links in this SCADA network are also redundant. Multiple communication paths to the 78 RTUs in the field are implemented by two PacBell telephone-wired networks (ADN and PSTN). In addition, there are several wireless links—a digital radio link, UHF radio link, and satellite links to each OCC. If one path fails there are two other alternatives.

Each node of Figure 10.5 is labeled with a number indicating its rank according to both connectivity and betweenness centralities. The two primary communication hubs—PacBell and GSTAR—are the most critical, with the Sunol node placing third. As you can see, links are not critical because of the redundancy of communication paths. If one fails, there are two other paths to take its place. Therefore, link betweenness ranks low relative to node betweenness and connectivity.

Redundant communication links reduce the likelihood of communication blackout, but it increases the network's spectral radius and therefore its self-organized criticality. This is particularly important when analyzing the spread of malicious software. Spectral radius of the SFPUC SCADA network is 3.67 versus mean connectivity of 3.25. The exceedence probability of a typical cascade caused by random injection of a malicious software exploit is also shown in Figure 10.5.

The exceedence probability distribution is very long tailed, so cascading malware quickly spreads to all nodes. Therefore, cyber exploits make this network fragile. Does this mean the entire SCADA network fails? The network is robust against physical threats, but much less robust against cyber exploits that take advantage of the percolated network. In this case, redundancy increased spectral radius, which increased risk of cascading failures while decreasing risk of physical failures. Once again, risk, resiliency, and robustness in complex CIKR may interact in counterintuitive ways.

10.8.2 Risk Reduction and Resource Allocation

The most significant hubs are the nodes with six and five links, respectively. These are the PacBell node with six links and the GSTAR satellite node with five links. Because of its unique radio link, the Sunol OCC node is the third most critical node in the SCADA network. As we shall see, failure of the PacBell, GSTAR, and Sunol hubs overwhelms the triple redundant network and causes the entire SCADA network to fail. But this is highly unlikely, because all three must fail at once.

Figure 10.6 shows the results of analyzing the impact of triple redundancy on risk and risk reduction for the three most critical components—nodes PacBell, GSTAR, and Sunol. At first glance it seems ridiculous to suggest that a satellite in space or the entire PacBell network system might be vulnerable to an attack. But if the threat comes from a cyber attack instead of a physical attack, it can disrupt satellites 23,000 miles in space as easily as placing a roadblock across an interstate highway. Similarly, the PacBell network

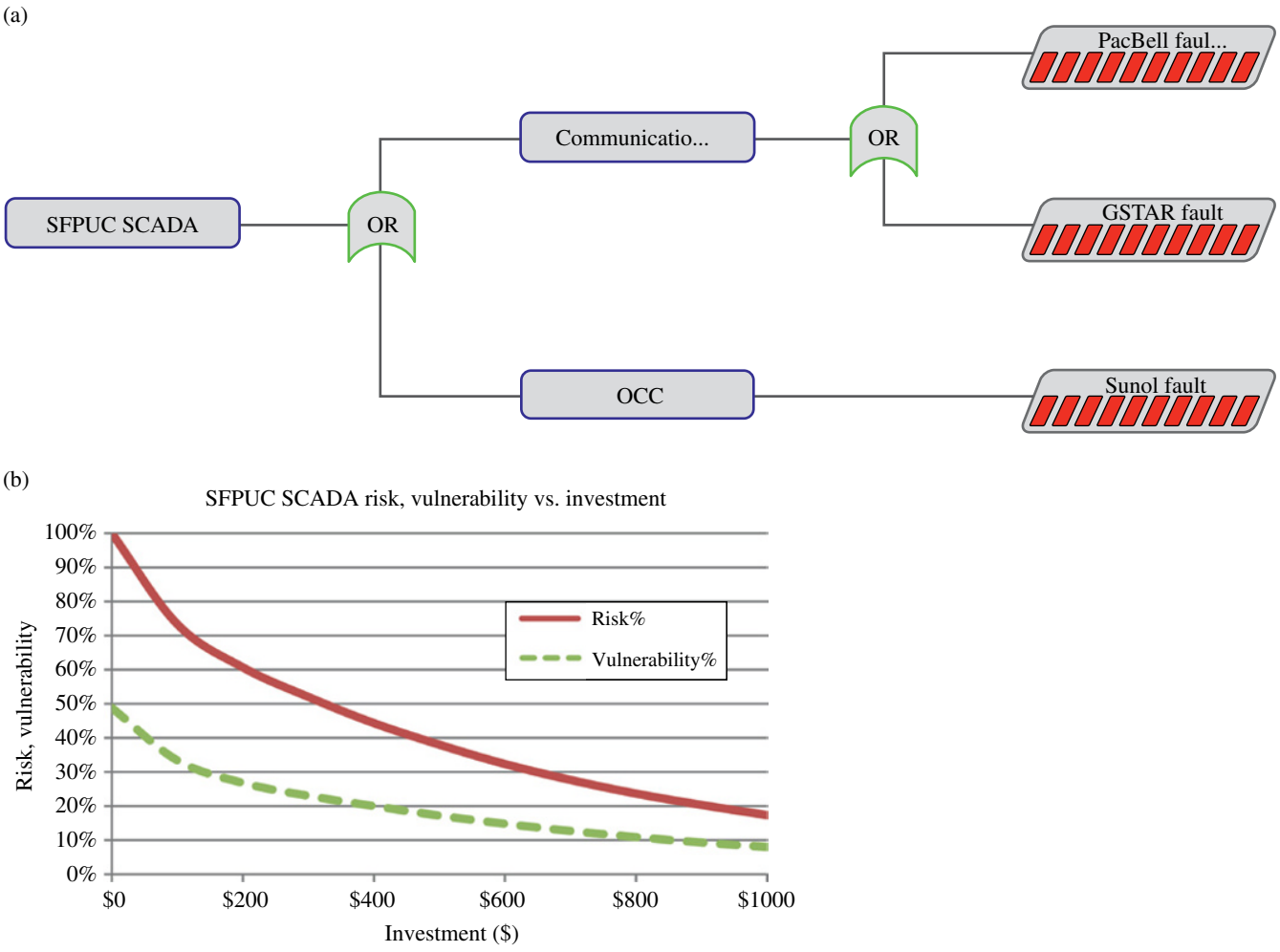


FIGURE 10.6 Fault tree of the three most critical nodes with hypothetical threat, vulnerability, consequence, and elimination cost estimates. (a) Fault tree containing the three most critical nodes: PacBell, GSTAR, and Sunol. (b) Risk and vulnerability reduction versus investment decline exponentially.

does not need to be physically destroyed to render it useless. Cyber attacks are extremely asymmetric because a single vandal or terrorist can launch a major denial of service attack that reduces the capacity or functionality of the entire PacBell communications network.

The question naturally arises, "what happens if these nodes fail?" How does individual vulnerability of each of these three threat–asset pairs affect the overall vulnerability of the entire network? Assuming the hypothetical values shown in Table 10.2, the risk and vulnerability reduction declines along an exponential curve as shown in Figure 10.6b. Furthermore, an investment of $600 thousand reduces risk from $300 thousand to $97 and vulnerability from 48.8 to 14.8%. This is a vast improvement over initial risk and vulnerability, but ROI is less than $1.00/$.

Note that resource allocation goes to the threat–asset pair with the highest return on investment—GSTAR. GSTAR consequence is highest, which means a small investment reduces risk more than a small investment in the other two threat–asset pairs. Maximizing ROI is distinctly different than allocating resources according to rank order. However, in this case optimal allocation is the same as rank ordering. In any case, diminishing returns eventually sets in, as shown in Figure 10.6b.

10.9 INDUSTRIAL CONTROL OF POWER PLANTS

The SFPUC SCADA network is an example of a simple ICS. Industrial control systems are extremely diverse, and one system does not represent all systems. The following analysis of a control system for power plants supplying power to an electrical power grid is explored to show the wide divergence of ICS architectures. It also is an opportunity to study the relationship among risk, resilience, and recovery time. Figure 10.7 shows the communication and control network for controlling a number of power plants connected to a regional power grid. The annotated network is labeled with cascade frequencies obtained by simulating 20,000 cascade failures.

10.9.1 Maximum PML

Figure 10.8 shows the results of node cascade simulations whereby a single node is selected at random followed by cascading that spreads with vulnerability V. The average value of V is 0.65, so spreading is extensive (see Fig. 10.7). A major hub acts as a super-spreader with cascade frequencies tapering off with distance from the hub. Also notice that the network is bimodal because it has a hub surrounded by adjacent nodes and a cluster of tightly connected nodes as shown in the lower left-hand corner of Figure 10.7. This bimodal topology will show up in the results.

Figure 10.8a shows the results of node cascading. Starting with a randomly selected node, adjacent nodes fail with probability V and spread the fault to their adjacent nodes. Eventually, the spreading dies out, or all nodes fail. The exceedence probability of failures of size x-axis or greater is plotted as a solid gray area in the top plot of Figure 10.8a, and the log–log plot of exceedence is plotted and fit to a straight line in the bottom graph. The slope of the straight line is the fractal dimension of the cascading network. Fractal dimension is 0.96 in Figure 10.8a and maximum PML risk is $876.7 thousand.

A budget of $500 thousand is allocated to nodes to reduce vulnerability. The result of cascading after vulnerability reduction is shown in Figure 10.8b. Note that fractal dimension increases to 2.17, indicating a more rapid drop in exceedence probability—the tail is shorter and/or not as fat—and suggesting a more resilient network. The maximum PML risk drops dramatically to $31.4 thousand—a 96% drop!

Optimal allocation of $500 thousand minimizes total network risk by maximizing ROI. In particular, more investment goes into vulnerability reduction of the most influential nodes, because we know that high influence and high connectivity equate with cascade frequency. The most influential nodes in Figure 10.7 are the hub, and the nodes surrounding the hub that are also highly connected. For example, static risk of the hub node was TVC = (1)(0.4)(500) = 200 before allocation and (1)(0.117)(500) = 58.64 after allocation.

10.9.2 Recovery

Vulnerability reduction has a significant impact on risk reduction. But it also has a significant impact on resilience and recovery time and effort. Recovery time depends on how many repair units operate in parallel and the mean time to repair a single node using a single repair unit. For example, a utility company may operate three repair trucks and crew at a time, but if six nodes require repair, they will be scheduled. As soon as one repair truck and crew

TABLE 10.2 Input and output values used in Figure 10.6 to evaluate risk reduction of the SFPUC SCADA network

Name	Threat (%)	Vulnerability (%)	Elimination cost ($)	Consequence ($)	Risk initial	Allocation ($)	Vulnerability reduced (%)	Risk reduced
PacBell fault	100.00	20.00	300.00	500.00	100.00	172.00	3.59	17.95
CSTAR fault	100.00	20.00	1500.00	750.00	150.00	257.14	11.97	89.76
Sunol fault	100.00	20.00	100.00	250.00	50.00	70.87	2.39	5.98

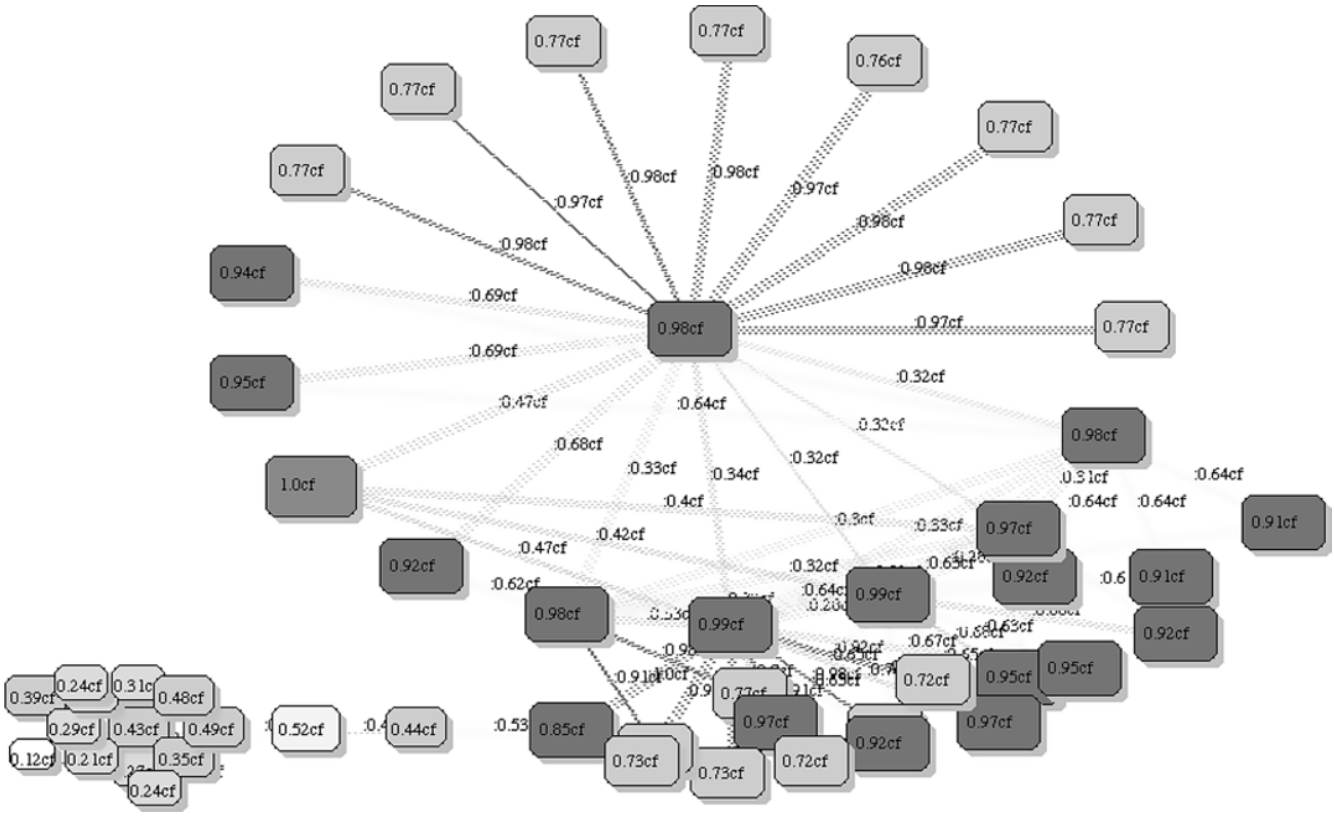


FIGURE 10.7 An industrial control system for control of power plants connected to a major power grid located in the Midwestern United States.

complete a repair, they move on to the next until all repairs have been done. We do not know the length of time to repair a single node, so we assume an average or mean repair time. In practice, actual repair times may vary, so the simulation of repair time is obtained by sampling from an exponential distribution.

Damage and repair times are estimated by simulating thousands of cascades. Figure 10.9 illustrates results obtained from 20,000 cascades. Before vulnerability reduction, mean damages total 1047.0, and mean recovery time is 55.5 time units assuming three repair units working in parallel, each requiring 10 time units on average to repair a node. After allocation of $500 thousand to reduce vulnerability, these numbers drop to 84.6 and 19.0, respectively. Damage declines by 92% and mean recovery time declines by 66%.

The top plot of Figure 10.9a appears quite different than the top plot of Figure 10.9b. Both graphs plot damage versus recovery time, but Figure 10.9a reveals the structure of the network—damages fall into two nearly separate clusters. In Figure 10.9a one cluster signifies low-damage, quick repair nodes, and the other cluster signifies high-damage, slow repairs. This is due to the bimodal topology of the network. It shows that cascading nodes are restricted to their neighborhoods.

On the other hand, Figure 10.9b shows less clustering in the damage versus repair time plot. Damage and repair time are more evenly distributed across the entire plot. This is due to the nature of the optimal resource allocation that maximized investment in ROI. Vulnerability is reduced such that ROI is flattened out, spreading risk over all nodes. High-risk nodes are reduced to low-risk nodes by optimal allocation.

10.9.3 Node Resilience

Figure 10.10 shows the results of resilience analysis. Recall that cascade resilience depends on two factors—vulnerability and spectral radius. In this case, spectral radius does not change from one simulation to the next; hence resilience depends solely on vulnerability. Figure 10.10a shows node cascade resilience before risk reduction, while Figure 10.10b shows node resilience after risk reduction by reducing vulnerability.

Figure 10.10a shows the original network of Figure 10.7 is not resilient, because it falls into the darkest color of the resilience chart. Recall that the resilience chart is the result of simulating thousands of cascades for each value of vulnerability, ranging from near zero to near 1.0, and noting how fractal dimension changes as vulnerability changes. This relationship falls on a straight line as shown in the figures. However, the vertical axis declines from a positive value to a negative value as vulnerability increases. This signifies lack of resilience.

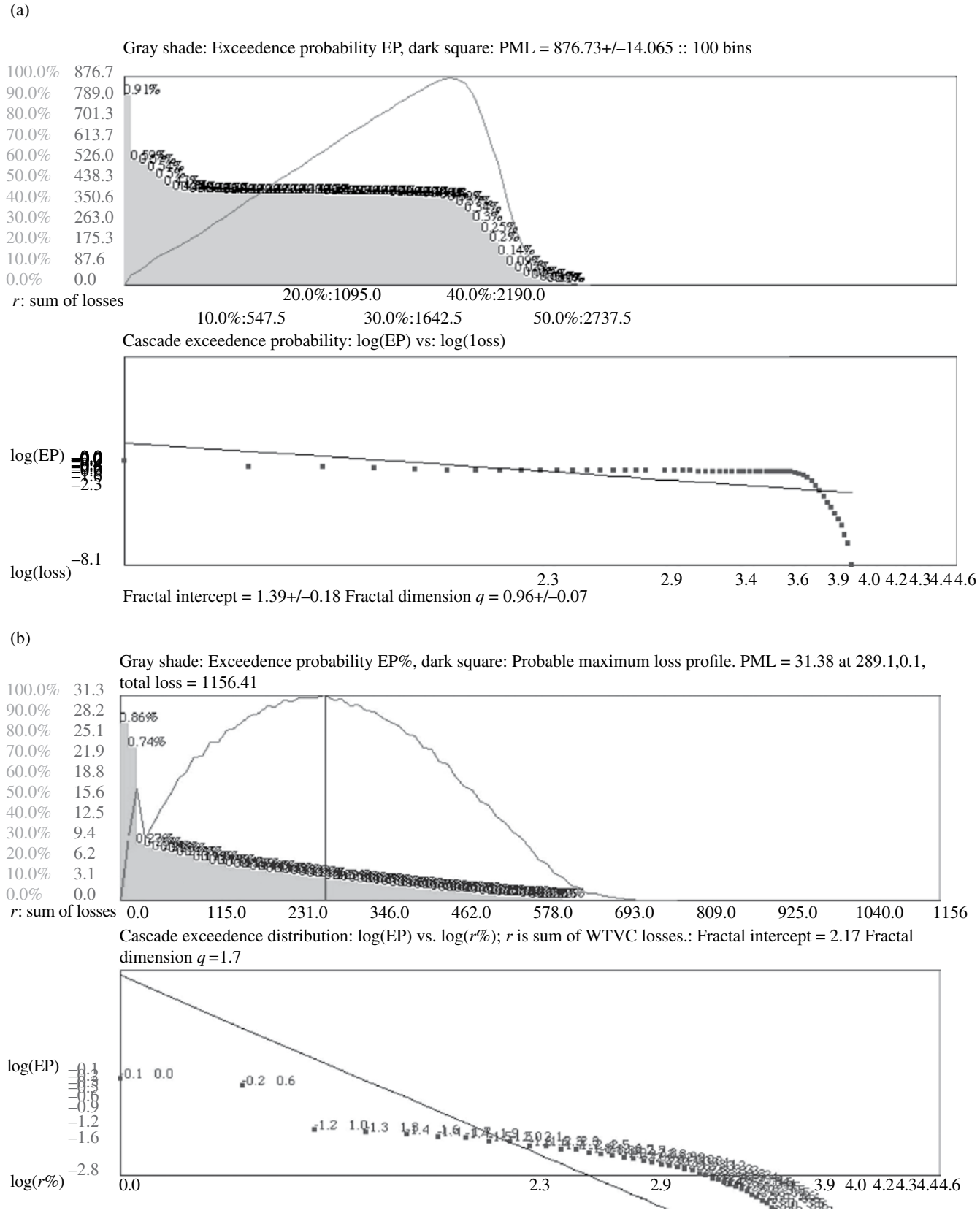


FIGURE 10.8 PML risk due to cascading before and after optimal allocation of $500 thousand to reduce vulnerability, V. (a) Maximum PML risk and exceedence probability before investment is $876.7 thousand with fractal dimension of 0.96. (b) Maximum PML risk and exceedence probability is $31.4 thousand with fractal dimension of 2.17 after investment and optimal allocation.

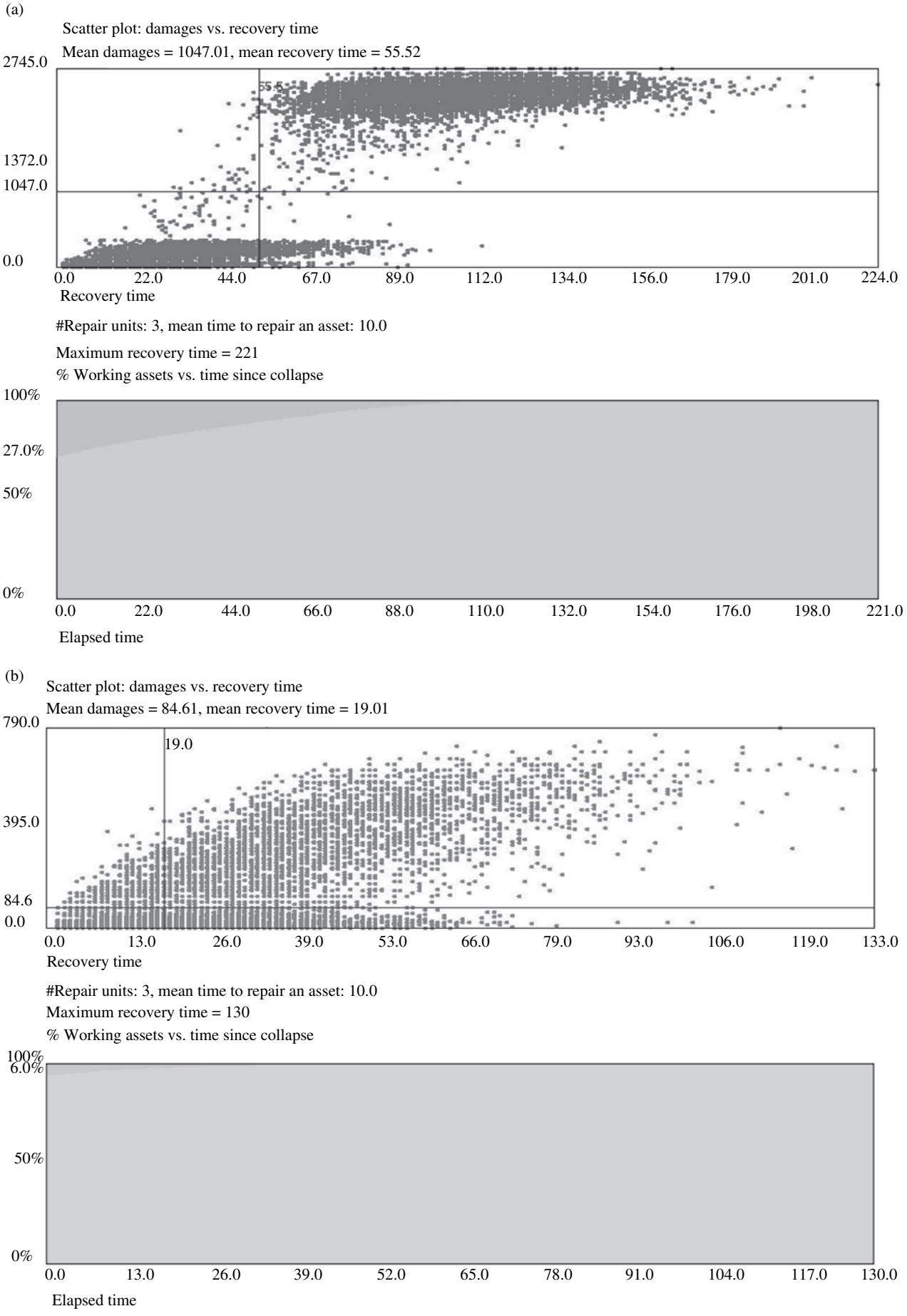


FIGURE 10.9 Top plot is damage versus recovery time. Bottom plot is percent of working assets versus elapsed time following a cascade collapse. Three repair units working in parallel, each with mean repair time of 10. (a) Recovery before vulnerability reduction: mean damage = 1047, mean recovery time = 221. (b). Recovery after vulnerability reduction: mean damage = 84.6, mean recovery time = 19.0.

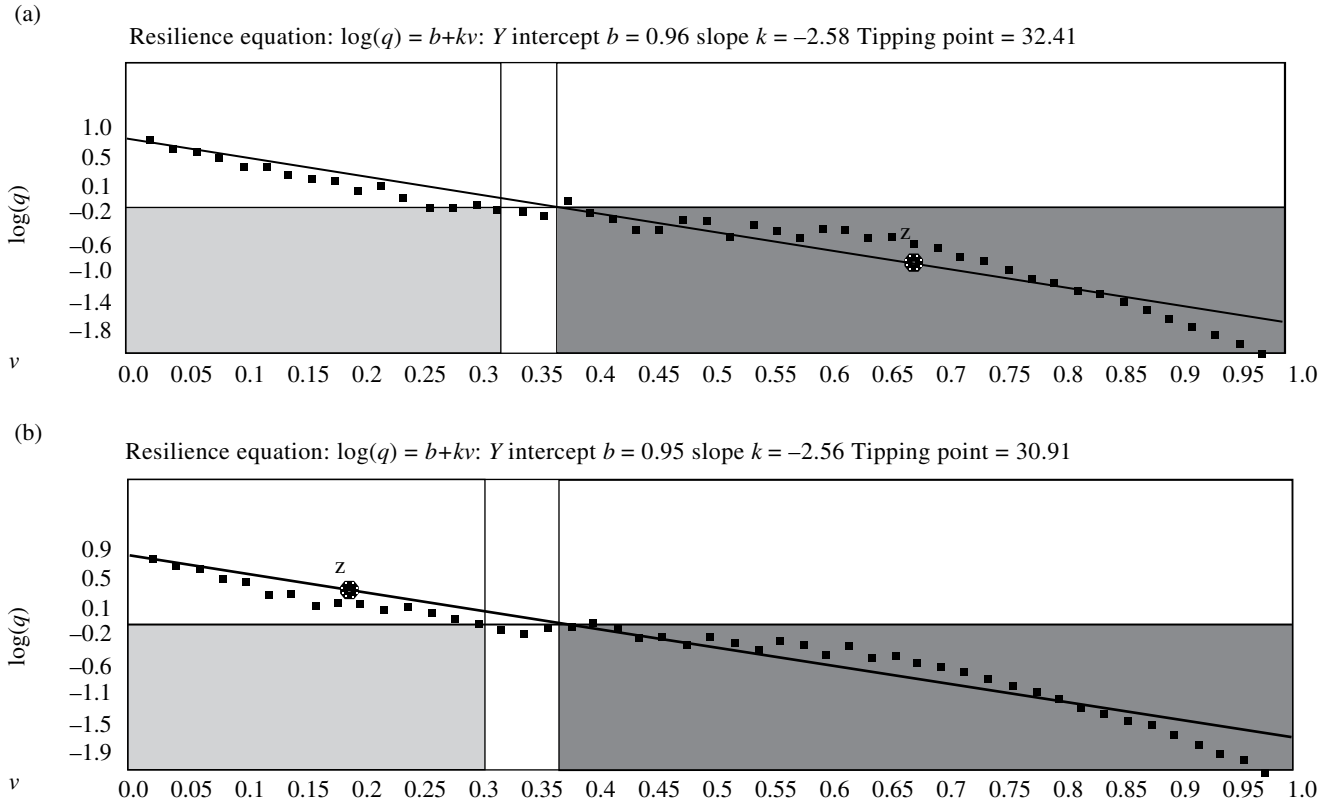


FIGURE 10.10 Top plot is node cascade resilience before an investment of $500 thousand to reduce vulnerability. Bottom plot is node cascade resilience after an investment to reduce vulnerability. Zones: shaded (safe), light (tipping point), and dark (unsafe). (a) Node resilience falls into the dark zone (darkest color) due to high vulnerability of nodes in the original network. (b) Node resilience falls into the shaded zone (light color) after vulnerability reduction.

Recall the fundamental resilience line establishes a critical vulnerability point, or tipping point, where cascading transitions from mild to extreme. This point separates network cascade resilience into zones:

> ***Critical vulnerability****: A network's critical vulnerability point is the value of vulnerability to cascading where the fundamental resilience line crosses zero. The shaded zone exists below this point, and the dark zone exists above this point. A light zone may exist if the straight line and simulation data cross as different points.*

The plots of Figure 10.10 contain light zones (light colored) because the simulation data and straight line differ. This difference introduces an uncertainty or amount of error in the estimate of critical vulnerability. Regardless, the light zone represents a phase transition from mild cascade collapse to extreme cascading. It also indicates a shift from resilient to not resilient status of the network.

Before the investment to reduce vulnerability, the network falls at point Z in Figure 10.10a, corresponding with an average node vulnerability of 0.65. After vulnerability reduction, the network falls at point Z in Figure 10.10b, corresponding with an average node vulnerability of 0.33. This moves the network from the dark zone into the shaded zone.

10.10 ANALYSIS

The SFPUC water SCADA and the power plant ICS examples focused on risk and resilience analysis using quantitative tools. The following and Table 10.3 summarizes recommendations more generally:

- Cyber intrusion is a major risk to control systems. Protect SCADA by isolating it from the corporate network, encrypting its data, and enforcing validated passwords and biometrics. Isolation may be implemented through physical separation or through firewall machinery that logically separates corporate networks from control system networks.[18]

[18]A firewall is a computer that filters or blocks input and output data ports, thus restricting user access to a network or subnetwork.

- Redundancy can reduce fault probabilities by an order of magnitude. Wherever a critical node is found, duplicate its function with double or triple redundant "backups." Because redundancy multiplies fault probabilities—instead of being additive—it is an effective means of protecting any system. But redundancy can also give operators a false sense of security, especially when it comes to cyber security. Computer viruses and worms can simultaneously infect all redundant computers, rendering them all unusable. In some cases, redundancy magnifies the spread of malware, because it increases self-organization.
- Communication security is as important as OCC security. Treatment plants, pipelines, and control centers are important, but do not overlook the network that connects them together. Careful implementation of user authentication, network isolation, and redundant communication links (as illustrated by the SFPUC example) can reduce physical risk. The links of the power plant ICS network can be further analyzed to identify the kill chain leading up to protected assets.
- The SFPUC water SCADA system is highly secure, because of triple redundancy in servers, communication links, and OCCs. This greatly improved system is an example of how to correctly increase SCADA security and harden typical water SCADA networks against natural and man-made attacks. But cyber attacks may still succeed against such a redundant system unless properly isolated and administered.
- The power plant example is highly fragile, because of its self-organized topology (hub-and-spoke structure) and high vulnerability. Redundancy will not reduce its vulnerability to cyber attacks. However, it can be made more resilient by reducing node vulnerability. It can also be made more resilient by restructuring it to reduce spectral radius, but this analysis was not done here.

SCADA is found in almost every industrial process and most critical infrastructures. It is pervasive. The trend is for even more automation because it lowers costs and increases speed and efficiency. The sectors most impacted are level 1 infrastructures—water, power, energy, and telecommunications. In 2013, over 50% of ICS-SCADA cyber exploits targeted the power grid, and 17% targeted the critical manufacturing sector. So, nearly 80% of all SCADA exploits were concentrated in these two sectors. Nonetheless, SCADA and related control systems are employed in other critical infrastructures such as transportation, food, and agriculture. These may be targeted next.

Attacks on infrastructures that use any automated control system can be extremely asymmetric. An inexpensive cyber attack on an SCADA network can bring down the entire network. Fortunately, the likelihood of a successful attack on the triple redundant SFPUC network is extremely low. Furthermore, it is not clear that damages would be very large—not nearly as consequential as the estimates used here.[19] A successful attack on water SCADA may lead to contamination or destruction of equipment, but there is no recorded case of SCADA attacks, leading to mass casualties.

On the other hand, a cyber attack on the highly structured power plant control network described here is likely to succeed and spread to nearly all parts of the network. This is due to two factors: self-organization into a hub-and-spoke structure with relatively high spectral radius and the relatively high vulnerability of nodes. This vulnerability represents the probability of malware spreading from nodes to adjacent nodes. The network is highly contagious.

Responsibility for SCADA security is scattered across governmental and commercial bureaucracies. This is unfortunate but understandable as SCADA applications are scattered across hundreds of industries. There is no SCADA ISAC because SCADA and control system security is vertical industry specific. According to Weiss, "end-users won't share critical information with an ISAC that could act as a policeman and also with an organization they don't know or trust."[20]

At this time in history, damages done by cyber terrorists have been minor. There have been no deaths and the cost to the economy has been relatively low. SCADA components are vulnerable, but damages are so low that SCADA security has not gained much attention as a major threat. Will this change if a major event occurs? It is important for us to differentiate between vulnerabilities and risk. A very low-consequence threat associated with a low fault probability is of little interest. Interest may still be low if a system is highly vulnerable but financial risk is extremely low. When both vulnerability and risk are high, we should be extremely interested in target hardening.

Even though SCADA attacks have failed to gain much public attention, the potential for major damage to SCADA networks and indirectly to the economy still remains. Thus, SCADA and other control system policies should focus on hardening of targets against cyber intrusion. DOE has provided 21 steps for protecting SCADA in the power sector (see Table 10.3). These steps are general enough to apply to all SCADA systems. They are policies that every security-conscious organization should follow.

[19]In general we do not know the extent of damage that might be inflicted by a successful cyber attack on some SCADA systems.

[20]Personal communication with Joe Weiss, April 2005.

TABLE 10.3 Department of Energy's 21 steps to SCADA security are easy to follow

1. Identify all connections to SCADA networks
2. Disconnect unnecessary connections to the SCADA network
3. Evaluate and strengthen the security of any remaining connections to the SCADA network
4. Harden SCADA networks by removing or disabling unnecessary services
5. Do not rely on proprietary protocols to protect your system
6. Implement the security features provided by device and system vendors
7. Establish strong controls over any medium that is used as a backdoor into the SCADA network
8. Implement internal and external intrusion detection systems and establish 24-h-a-day incident monitoring
9. Perform technical audits of SCADA devices and networks, and any other connected network to identify security concerns
10. Conduct physical security surveys and assess all remote sites connected to the SCADA network to evaluate their security
11. Establish SCADA "Red Teams" to identify and evaluate possible attack scenarios
12. Clearly define cyber security roles, responsibilities, and authorities for managers, system administrators, and users
13. Document network architecture and identify systems that serve critical functions or contain sensitive information that requires additional levels of protection
14. Establish a rigorous, ongoing risk management process
15. Establish a network protection strategy based on the principle of defense-in-depth
16. Clearly identify cyber security requirements
17. Establish effective configuration management processes
18. Conduct routine self-assessments
19. Establish system backups and disaster recovery plans
20. Senior organizational leadership should establish expectations for cyber security performance and hold individuals accountable for their performance
21. Establish policies and conduct training to minimize the likelihood that organizational personnel will inadvertently disclose sensitive information regarding SCADA system design, operations, or security controls

Source: From http://www.utc.org/?v2_group=0&p=3629.

10.11 EXERCISES

1. What is SCADA?
   a. Secure communications for data analysis
   b. Secure communications for data acquisition
   c. Supervisory control and data analysis
   d. Supervisory control and data acquisition
   e. Supervisory control and distributed analysis
2. Why is encryption *not* used more often in SCADA communications?
   a. Keys are not long enough.
   b. Encryption introduces overhead and cost.
   c. Analog communications cannot be encrypted.
   d. Cellular modems can be war dialed.
   e. There is nothing secret in most water supply systems.
3. SCADA security can be improved by:
   a. Isolating SCADA networks from corporate networks
   b. Increasing latency of the network
   c. Increasing reliability of servers
   d. Reducing government regulation
   e. Increasing governmental regulation
4. Who is the lead federal agency responsible for protecting critical infrastructure from SCADA exploits?
   a. Department of Homeland Security
   b. EPA
   c. Department of Energy
   d. Department of Treasury
   e. NIST
5. What is the argument against redundancy?
   a. It does not always work
   b. It adds to the cost of a system
   c. It reduces fault probabilities
   d. It reduces financial risk
   e. It increases vulnerability to malicious software cascades
6. In the SFPUC SCADA case study, why did redundancy produce a lower sector risk?
   a. Redundancy increased percolation of the network.
   b. Consequence was reduced.
   c. Redundancy has a multiplicative effect on vulnerability.
   d. Fault trees are not perfect.
   e. All of the above.
7. Why is cyber intrusion such a major threat to SCADA?
   a. SCADA systems are notoriously open to cyber attacks.
   b. Cyber SCADA exploits have historically been disastrous.
   c. Cyber SCADA attacks have killed people.
   d. Cyber SCADA protection is expensive.
   e. Scientific studies have concluded these are the worst vulnerabilities.

8. CSET is a:
   a. Question-and-answer tool for assessing SCADA security
   b. A method of assessing cyber security risk
   c. A method of assessing cyber security vulnerability
   d. An acronym for Computer Security Evaluation Test
   e. None of the above
9. SCADA is used in:
   a. Power grid monitoring and control
   b. Water system monitoring and control
   c. Gas and oil pipeline monitoring and control
   d. Transportation system monitoring and control
   e. All of the above
10. Which of the following make SCADA different than general IT systems?
    a. Loss of life
    b. Loss of hardware
    c. Loss of information
    d. Human safety
    e. Production delays

10.12 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. Why are there so many different risk and resilience frameworks for SCADA and ICS?

B. Compare the lack of cybersecurity in the design of the Internet with lack of cybersecurity of most SCADA systems. Why are they both fragile with respect to cyber security even though they were designed and deployed by two entirely separate industries?

C. Cyber attacks against infrastructure systems have steadily increased over the years. Why, and by whom?

D. Explain in your own words why redundancy may lead to less resilience against cascading in an SCADA network. Recall the Paradox of Redundancy and the Braess paradox.

E. Is cybersecurity more critical to SCADA than physical security? Explain and justify your answer.

11

WATER AND WATER TREATMENT

PPD-21 defines the security goal of the water and wastewater treatment sector as "… a secure and resilient drinking water and wastewater infrastructure that provides clean and safe water as an integral part of daily life, ensuring the economic vitality of and public confidence in the Nation's drinking water and wastewater service through a layered defense of effective preparedness and security practices in the sector." Note that the emphasis is on drinking water, which is approximately 15% of the county's water supply chain. Agricultural and industrial water is excluded from this sector. Also note that water, in the form of hydroelectric power, is also connected to other CIKR sectors such as power and transportation.

This chapter traces the evolution of water as a valuable resource protected by legislation and regulation to a critical infrastructure that supplies drinking water to 300 million Americans and is indirectly linked to food production (food/agriculture infrastructure) as well as industrial production capacity. It then analyzes one of the nation's largest water systems—the Hetch Hetchy network that supplies water to the San Francisco Bay Area. This case study once again illustrates risk assessment and shows how to optimally allocate water supply improvement funds for the protection and response to this CIKR's hazards.

In addition, this chapter traces water supply legislation and shows how it has evolved from a public health issue (biological contamination of drinking water) to an environmental protection issue (chemical and radiological contamination), to bioterrorism, and finally to climate change and neglect. At one time the main concern was that terrorists might disrupt the supply of drinking water with a *denial-of-service* (DoS) attack, biological contamination attack, or bombing. This concern has shifted to environmental and political neglect as America's water supply has deteriorated. Environmental conditions may reduce the availability of water to the millions of people who depend on it and impact other sectors through flooding power plants/grids or transportation systems. To illustrate these potential threats and propose strategies for protection and response, the San Francisco water system is analyzed against biological treatment plant, earthquake pipeline, and Supervisory Control and Data Acquisition (SCADA) treatment plant threat–asset pairs using hypothetical data.

Finally, this chapter compares the RAMCAP™ framework recommended by the American Society of Mechanical Engineers (ASME) with fault tree analysis.[1] The Department of Homeland Security recommends RAMCAP™ be used for risk analysis of the water CIKR. Both RAMCAP™ and fault tree analysis use the familiar TVC formula for risk, but model-based risk analysis (MBRA) fault trees may be used for resource allocation. RAMCAP™ is an alternative risk assessment to MBRA's network and fault tree analysis tools.

In addition, the American Water Works Association (AWWA) Water Utility Council initiated a project to address cyber attacks on water industrial control systems in February 2013. The AWWA developed a cybersecurity guidance tool is a voluntary, sector-specific approach for adopting the

[1]ASME Innovative Technologies Institute, LLC 1828 L Street, NW Suite 906 Washington, DC, 20036 info@asme-iti.org (202) 785-7388, www.asme-iti.org

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

NIST-CSF (Cybersecurity Framework) as expressed by the Water Sector Coordinating Council.

The following major topics and concepts are described in detail:

- *Purity versus terrorism*: Public health legislation at the turn of the twentieth century was focused on water purity and the prevention of disease. The US Public Health Service (USPHS) was responsible for protecting drinking water in communities across the country. Today the US Environmental Protection Agency (US EPA) has responsibility for protecting the water supply system from biological, chemical, and radiological contamination as well as countering DoS attacks perpetrated by terrorists.
- *Drinking versus agricultural and industrial uses of water*: By far the preponderance of legislation and regulation of water has been aimed at drinking water, and yet over 80% of the water supply is used for agricultural and industrial applications. Intelligent life on this planet depends on water, but so does civilized food production and industrial economy.
- *Safe Drinking Water Act (SDWA) of 1974*: The SDWA of 1974 is the foundation upon which modern water regulation is based. It also transferred responsibility from the US Public Health Department to the US EPA. The SDWA has been modified many times since 1974, but it still stands as the foundation for contemporary water safety.
- *Bioterrorism Act of 2002*: Title IV of the *Public Health Security and Bioterrorism Preparedness and Response Act of 2002* extended the SDWA of 1974 to include a new threat: terrorism. It also directs water communities to perform vulnerability analysis for a new failure mode—DoS (cutting off the supply of water entirely).
- *New threats*: Climate change and neglect are the new threats to drinking water as the United States becomes dryer in some places, flooding occurs in other places, and water infrastructure begins to decay everywhere due to aging and neglect.
- *Case study*: The massive and vital Hetch Hetchy water and power supply network of the San Francisco Bay Area is shown to be critical due to the interdependencies among water, power, San Francisco International Airport, Silicon Valley computer industry, and surrounding metropolitan communities that depend on the Hetch Hetchy water supply. While its spectral radius is relatively low, its betweenness centrality is relatively high. Thus it is resilient against cascade failures, but fragile against flow disruptions. Critical assets are the major pipelines, treatment plants, power plants, and large reservoirs.
- *MBRA*: MBRA uses hypothetical data to illustrate how best to allocate risk reduction funds by protecting reservoirs, treatment plants, pipes, and powerhouses against bombings, earthquakes, power outages, corrosion, and chem/bio attacks. The most critical assets in the San Francisco water system lie along a critical path defined by high betweenness. The most critical threat–asset pairs are shown to be high-consequence pairs threatened by earthquakes. Yet, sector resilience is modestly high because of many alternative sources and paths from sources to destinations.
- *Risk methods*: RAMCAP™ is the recommended risk assessment tool for the water sector. It is based on the PRA equation: R = TVC and uses a risk ranking strategy to allocate resources. In comparison, MBRA extends R = TVC by including fault tree logic and optimal resource allocation that minimizes risk. MBRA of Hetch Hetchy (using hypothetical data) minimizes risk by allocating most funding toward earthquake retrofitting assets along the critical path defined by the highest-betweenness nodes and links of the Hetch Hetchy water and power network.

11.1 FROM GERMS TO TERRORISTS

Prior to the development of the germ theory by Louis Pasteur in the 1880s, water was simply a resource to be exploited for powering water wheels and quenching the thirst of humans, animals, and crops. But soon after Pasteur developed his theory, water became a recognized vector for the transmission of diseases. Dr. John Snow showed how cholera was transmitted from wells to homes via water in 1885 [1]. And by 1914 the USPHS began setting standards for the *bacteriological* purity of drinking water. But these standards had to be promulgated by water utilities that served rather sizeable communities. Utilities maximized profit—sometimes at the expense of water purity—and small utilities were not closely monitored. As a consequence, it took the country decades to "purify" the drinking water consumed in the United States.

As more and more unhealthy substances were shown to exist in drinking water, they were added to the list of substances regulated by the *maximum contaminant level* (MCL) standard. Bacteriological standards were revised in 1925, in 1946, and again in1962. In 1960 a USPHS study showed that only 60% of drinking water met PHS purity standards. Consequently, a 1962 revision increased the number of substances falling within the regulations to 28 substances—the most rigorous standards until 1974. Table 11.1 lists these substances.

A 1972 USPHS study of the Mississippi River found 36 *chemical* contaminants remaining in drinking water *after* treatment plants had processed it. The treatment plants were not filtering out these hazardous chemicals, pollution, pesticides, and other chemical and radiological contaminants. Biological contamination was but one of many contaminants in the Mississippi River water supply. Chemicals had become a bigger problem. The 1972 study underscored the

TABLE 11.1 These contaminants are regulated per the 1962 Public Health Service standards

Alkyl benzene sulfonate (ABS)
Arsenic
Barium
Beta and photon emitters
Cadmium
Carbon chloroform extract (CCE)
Chloride
Chromium
Color
Copper
Cyanide
Fluoride
Gross alpha emitters
Iron
Lead
Manganese
Nitrate
Phenols
Radium-226
Selenium
Silver
Strontium-90
Sulfate
Threshold odor number
Total coliform
Total dissolved solids
Turbidity
Zinc

importance of filtering out nonbiological contamination and led to the creation of the modern foundation of water legislation—the *SDWA of 1974*.

11.1.1 Safe Drinking Water Act

The SDWA establishes the foundation of modern regulations for protecting the purity of water *and water systems*. Because the list of contaminants had grown to include chemicals and other hazardous materials, the responsibility for protecting drinking water and associated processing systems was transferred to the US EPA. This foundational act was revised in 1986 and again in1996 and 2002, but it remains the bedrock of water legislation.

The concept of water as a critical resource expanded once again as acts of terrorism multiplied during the 1990s. On May 22, 1998, President Clinton signed Presidential Decision Directive 63 (PDD-63), which identified, among other sectors, *drinking* water as one of America's critical infrastructures. People cannot survive longer without food than without water. Terrorists merely need to deny water service for a few days or week to cause major disruptions in the health, environment, and commerce of the country.

By 2002, water "purity" legislation had evolved from biological to environmental and then to DoS "contamination." PDD-63 identified the issue, but the Bioterrorism Act added terrorism to the list of contaminants. The *Public Health Security and Bioterrorism Preparedness and Response Act of 2002* was signed into law by President George W. Bush on June 12, 2002. It was the most significant event affecting water security since the SDWA of 1974. Title IV of this act addresses the water sector and provides a number of penalties for perpetrators of attacks on water systems.

Shortly after the Bioterrorism Act of 2002 was signed, the US EPA completed the first *classified* Baseline Threat Report describing likely modes of terrorist attack and outlining the parameters for vulnerability assessments by *community* water systems. This report remains classified. One can only speculate that the threats identified by the US EPA report are similar to the ones identified in the case study described later in Section 11.5.

11.1.2 The WaterISAC

In December 2002 the US EPA provided funds to the AWWA—a professional society for water system professionals—to form the Water Information Sharing and Analysis Center (WaterISAC) as prescribed by the National Strategy for Critical Infrastructure Protection. The WaterISAC is a consortium of professional associations and vendors focused on the promotion of water works safety and security. It brings together the private and public sector to implement the strategies of the SDWA and its descendants. It also provides training and education to its members in subjects such as vulnerability analysis and risk assessment.

The WaterISAC Board of Managers is composed of water utility managers appointed by the national drinking water and wastewater organizations below. There are also two at-large seats, filled by the Board of Managers. Typical members are:

American Water Works Association
Association of Metropolitan Sewerage Agencies
Association of Metropolitan Water Agencies
AWWA Research Foundation
National Association of Water Companies
National Rural Water Association
Water Environment Federation
Water Environment Research Foundation

The WaterISAC is a bridge between the public and private sectors operating within the water sector. It has established the following goals and provides the following products for its members:

- Alerts on potential terrorist activity.
- Aggregation of information on water security from federal homeland security, intelligence, law enforcement, public health, and environment agencies.

- Maintain databases of chemical, biological, and radiological (CBR) agents.
- Identify physical vulnerabilities and security solutions.
- Provide its members with notification of cyber vulnerabilities and technical fixes.
- Perform research and publish reports and other information.
- Provide a secure means for reporting security incidents.
- Recommend/provide vulnerability assessment tools and resources.
- Provide emergency preparedness and response resources.
- Provide secure electronic bulletin boards and chat rooms on security topics.
- Summarize open-source security information.

11.2 FOUNDATIONS: SDWA OF 1974

The SDWA of 1974 assigns responsibility for water safety to the US EPA. But the focus prior to 1974 was on biological purity. After 1974 the focus expanded to CBR purity. The shift from biological contamination to environmental pollutants signaled a phase shift in public policy regarding water. One more shift in policy direction occurred in 2002 when bioterrorism was added to the SDWA foundation.

Minor modifications in 1996 broadened the scope of responsibility of the US EPA. See Table 11.2. The EPA now has responsibility for entire water supply systems—not just drinking water coming from household taps, but also the US EPA is responsible for protecting the entire system including water from rivers, lakes, pipes, and treatment plants. This includes protection against physical, biological, chemical, radiological, and cyber threats. However, there are some exceptions to this policy.

The regulatory power of the US EPA does not extend to all water systems. The regulation distinguishes *community* water supply systems from *private drinking* water systems and other systems such as agricultural and industrial water. For example, it does not regulate private wells serving 25 or fewer consumers. Public or *community water* systems must serve at least 3300 consumers to fall within the EPA's jurisdiction. In addition, the stringency of the law increases as the size of the water supply system increases.

The US EPA does not assume full responsibility for enforcing the SDWA. Instead its strategy is to partner with states, tribes, and private utilities. It aims to regulate and fund local enforcement of the *National Primary Drinking Water Regulations*, which define enforceable MCL for particular contaminants in drinking water (see Table 11.3). It specifies that certain proven methods of decontamination be used to treat water to remove contaminants. It also sets standards for drinking water communities according to the number of people served by the system.

TABLE 11.2 The 1996 SDWA amendments require US EPA to enforce the following

Consumer confidence reports	All community water systems must prepare and distribute an annual report about the water they provide, including information on detected contaminants, possible health effects, and the water's source
Cost–benefit analysis	US EPA must conduct a thorough cost–benefit analysis for every new standard to determine whether the benefits of a drinking water standard justify the costs
Drinking water state revolving fund	States can use this fund to help water systems make infrastructure or management improvements or to help systems assess and protect their source water
Microbial contaminants and disinfection by-products	US EPA is required to strengthen protection for microbial contaminants, including cryptosporidium while strengthening control over the by-products of chemical disinfection. Two new drinking water rules in November 1998 addressed these issues; others will follow
Operator certification	Water system operators must be certified to ensure that systems are operated safely. US EPA issued guidelines in February 1999 specifying minimum standards for the certification and recertification of the operators of community and nontransient, noncommunity water systems
Public information and consultation	SDWA emphasizes that consumers have a right to know what is in their drinking water, where it comes from, how it is treated, and how to help protect it. US EPA distributes public information materials (through its Safe Drinking Water Hotline, Safe Water Web Site, and Water Resource Center) and holds public meetings, working with states, tribes, water systems, and environmental and civic groups, to encourage public involvement
Small water systems	Small water systems are given special consideration and resources under SDWA to make sure they have the managerial, financial, and technical ability to comply with drinking water standards
Source water assessment programs	Every state must conduct an assessment of its sources of drinking water (rivers, lakes, reservoirs, springs, and groundwater wells) to identify significant potential sources of contamination and to determine how susceptible the sources are to these threats

Understanding the Safe Drinking Water Act, December 1999, EPA 810-F-99-008. US-EPA's Office of Ground Water and Drinking Water web site: http://www.epa.gov/safewater/

11.3 THE BIOTERRORISM ACT OF 2002

Title IV of the Bioterrorism Act of 2002 extends the SDWA to cover terrorism and modern asymmetric threats such as SCADA attacks and insider attacks from employees of water

TABLE 11.3 Input data for the top five assets in the MBRA network model of Hetch Hetchy ranked according to consequences shows large differences in elimination costs

Name	Threat (%)	Vulnerability (%)	Consequence $(millions)	Prevention cost $(millions)	Response cost $(millions)	Risk initial $(millions)	Risk reduced $(millions)
Sunol Water Treatment	50.00	100.00	1000.00	20.00	100.00	500.00	9.90
Kirkwood Powerhouse	50.00	100.00	1500.00	10.00	25.00	750.00	12.96
Holm Powerhouse	50.00	100.00	1500.00	5.00	1.00	750.00	7.42
New Don Pedro Reservoir	50.00	100.00	1000.00	1000.00	10.00	500.00	5.00
Tracy Water Treatment	50.00	100.00	1000.00	25.00	25.00	500.00	9.81

treatment plants.[2] Water SCADA includes the computer and digital network infrastructure that supports the surveillance and operation of water, power, and energy sectors.

The 2002 act recommends hardening of targets by adding intruder detection equipment, installing fences, gating, lighting, locks, tamper-proof hydrants, and making improvements to ICS-SCADA hardware and software. It provides funds for training in operations and the handling of chemicals. It requires that water works employees and contractors submit to security screening and provides penalties for breach of confidentiality.

Some highlights of the act are as follows:

- Provided up to $160M in FY02, "such sums as may be necessary" in FY03-FY05 to (1) perform physical and SCADA vulnerability analysis of all systems with 3300 or more consumers according to the following timetable:

  March 2003 for communities of 100,000 or more.

  December 2003 for communities of 50,000–100,000.

  June 2004 for communities of 3,300–50,000 consumers.
- Restricts who has access to vulnerability assessment information and specifies penalties of up to 1 year in prison for anyone who "recklessly reveals such assessments." The results of RAMCAP™ risk assessments on water systems are confidential.
- Grants up to $5M for small communities (<3300 consumers).
- Requires all communities to develop an emergency response plan to "obviate or significantly lessen impact of terrorist attacks."
- Provides up to $15M in FY02, "such sums as may be necessary" in FY03-FY05 to:

  Work with the Centers for Disease Control (CDC) to "prevent, detect, and respond" to chemical, biological, and radioactive contamination of water.

  Review methods by which terrorists can disrupt supply or safety.

  Review means of providing alternative supply in event of disruption.

  Create a WaterISAC.
- Amend SDWA to extend wording about *water safety* to include wording about disruption of services by *terrorists*.

[2]http://www.fda.gov/oc/bioterrorism/PL107-188.html#title4

11.3.1 Is Water for Drinking?

After a century of focusing on biological, then environmental, and now terrorist threats and climate change to the water supply, the US EPA is working with states, tribes, drinking water and wastewater utilities (water utilities), and other partners to enhance the security of water, water works, sources of water, and wastewater utilities. It has set the following objectives for itself:

1. EPA will work with the states, tribes, drinking water and wastewater utilities (water utilities), and other partners to enhance the security of water and wastewater utilities.
2. EPA will work with the states, tribes, and other partners to enhance security in the chemical and oil industry.
3. EPA will work with other federal agencies, the building industry, and other partners to help reduce the vulnerability of indoor environments in buildings to CBR incidents.
4. EPA will help to ensure that critical environmental threat monitoring information and technologies are available to the private sector, federal counterparts, and state and local government to assist in threat detection.
5. EPA will be an active participant in national security and homeland security efforts pertaining to food, transportation, and energy.
6. EPA will manage its federal, civil, and criminal enforcement programs to meet our homeland security, counterterrorism, and antiterrorism responsibilities under Presidential Decision Directives (PDD) 39, 62, and 63 and environmental civil and criminal statutes.

But the national strategy as implemented by the US EPA addresses only a portion of the problem. In California, for example, 80% of the water managed via supply systems, treatment plants, aqueducts, and regulated utilities goes to agriculture, not drinking water. And this does not address the needs of industry. Water is needed to process silicon into computer chips in the $370 billion semiconductor industry. Without water, Silicon Valley would shrivel up as quickly as the Central Valley (major agricultural area of California). In addition, major hydroelectric power plants depend on the abundance of water to generate power for the San Francisco International Airport, for example. The famous Hetch Hetchy water supply system in Northern California provides water and power to 2.4 million inhabitants in the San Francisco Bay Area, but it also powers the San Francisco International Airport as well as itself. Without water there is no power and without power there is no water.

Thus the question is, "should the water sector be extended beyond drinking water?" Agricultural and industrial uses of water have become as important to national security as drinking water, so why not incorporate these interdependencies? These questions suggest that the concept of DoS—as applied to all water supply systems—is a major vulnerability to public health, agriculture, and the industrial base. Water directly affects at least three of the critical infrastructures defined in the National Strategy and indirectly affects the other sectors.

The following case study illustrates the interdependency of water with the economy and livability of the San Francisco Bay Area. It underscores the vulnerability of a water supply that quenches the thirst of a major metropolitan area—one that serves the famous Silicon Valley, perhaps America's most powerful generator of economic power, and is a "cousin" to the California Aqueduct system that supports one of the largest and most productive agricultural regions of the United States.

11.3.2 Climate Change and Rot: The New Threats

Since 9/11 water security rapidly changed phases from concern for biological purity to bioterrorism and, more recently, to climate change and the availability of water. Climatology and ecology is a vast topic outside of the scope of this textbook. But climate change will impact the water CIKR sector dramatically over the following decades. Some of the facts pertaining to water and climate change are summarized here[3]:

- While annual precipitation over the continental United States has increased by two inches between 1895 and 2011, the distribution of rainfall is shifting from south to north. The southwest is especially subject to declines in rainfall.
- Precipitation events are becoming more intense, lengthening the long-tailed distribution of exceedence versus intensity. Number and intensity of the heaviest precipitation events (storms) will increase everywhere in the United States.
- Conversely, dry spells are projected to increase in length in most regions especially in the south and southwest.
- There is an increased risk of flooding in many parts of the United States, for example, Northwest and Midwest, decreasing in the southwest and southeast.
- Total freshwater withdrawals and consumptive uses have leveled off nationally since 1980 at 350 billion gallons of withdrawn water and 100 billion gallons of consumptive water per day despite the addition of 68 million people from 1980 to 2005. Irrigation and all electric power plant cooling withdrawals account for approximately 77% of total withdrawals, municipal and industrial for 20%, and livestock and aquaculture for 3%. Most thermoelectric withdrawals are returned back to rivers after cooling, while most irrigation withdrawals are consumed by the processes of evapotranspiration and plant growth. Thus, consumptive water use is dominated by irrigation (81%) followed distantly by municipal and industrial (8%) and the remaining water uses (5%).
- Major threats to the water supply are drought, flooding, and worn-out drinking water systems (pipes, pumps, treatment plants). The American Society of Civil Engineers (ASCE) gave water infrastructure a grade of D in 2017.[4]

The last item above is of immediate concern. According to the ASCE,

> Drinking water is delivered via one million miles of pipes across the country. Many of those pipes were laid in the early to mid-20th century with a lifespan of 75 to 100 years. The quality of drinking water in the United States remains high, but legacy and emerging contaminants continue to require close attention. While water consumption is down, there are still an estimated 240,000 water main breaks per year in the United States, wasting over two trillion gallons of treated drinking water. According to the American Water Works Association, an estimated $1 trillion is necessary to maintain and expand service to meet demands over the next 25 years.[5]

The ASCE has been campaigning for $105 billion in renewal funding for the past decade, but the funding has fallen short. "In 2014, Congress authorized a new mechanism to fund primarily large water infrastructure projects over $20 million through the Water Infrastructure Finance and Innovation Act (WIFIA). In 2016 Congress appropriated $17 million in funds for the program. It is estimated that using WIFIA's full financial leveraging ability that a single dollar injected into

[3]https://nca2014.globalchange.gov/report/sectors/water

[4]https://www.infrastructurereportcard.org/

[5]https://www.infrastructurereportcard.org/cat-item/drinking-water/

the program can create $50 dollars for project lending. Under current appropriations, EPA estimates that current budget authority may provide more than $1 billion in credit assistance and may finance over $2 billion in water infrastructure investment."

11.4 THE ARCHITECTURE OF WATER SYSTEMS

Community water systems serve 3300 or more consumers. They are typically vertically integrated monopolies that own and operate all aspects of the water supply for a community (see Fig. 11.1). For example, the City and County of San Francisco owns and operates the San Francisco Public Utilities Commission (SFPUC) that provides water, wastewater, and electric power services to San Francisco and surrounding cities such as Alameda, San Mateo, and Santa Clara; Muni (public transportation); and San Francisco International Airport.

The SFPUC manages the collection of water from rivers and lakes; treatment facilities such as the Tracy and Sunol treatment plants studied in Chapter 10; storage in the form of temples and reservoirs; and distribution through a network of tunnels and pipelines. It owns and operates utility trucks and fire apparatus to protect its watershed located in the *Hetch Hetchy* region near Yellowstone Park in the Sierra Mountains—175 miles east of the Bay Area—and is responsible for treating wastewater before discharging it into the San Francisco Bay and the Pacific Ocean.[6]

In 1997, the city of Seattle, Washington, established Seattle Public Utilities (SPU) to provide water, sewer, drainage, and garbage services for 1.3 million people in King County, Washington. Similarly, District of Columbia Water and Sewer Authority (DC WASA) owns and operates the Washington, DC, water system including supplying drinking water, wastewater treatment, and fire hydrants since 1996, and the New York City Municipal Water Finance Authority is a public benefit corporation established by the New York City Municipal Water Finance Authority Act of 1984.

These vertical monopolies typically manage all levels shown in Figure 11.1:

- Collection of runoff from rivers and lakes in watershed territories such as the Hetch Hetchy lake area.
- Treatment of water prior to distribution to consumers such as the Sunol treatment facility.
- Storage of water in reservoirs or storage temples such as the 4-day supply held in storage tanks called temples, around the city of San Francisco as well as reservoirs outside of the city.

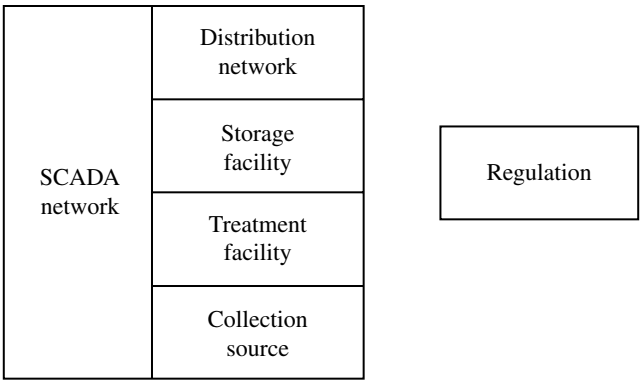


FIGURE 11.1 Typical community water systems are vertically integrated natural monopolies as shown here.

- Distribution of water through pipes, tunnels, and rivers such as the Crystal Springs Tunnel south of San Francisco.
- Monitoring and control of the entire system is handled by an ICS-SCADA network as described in Chapter 10.
- Legislation provides the monopoly's authority as well as the chemical, biological, and counterterrorism regulation as dictated by the federal government.

11.4.1 The Law of the River

Not all water systems are regulated at the local level. For example, the Colorado River supplies water to 27 million consumers in 7 states and 2 countries. Under a contentious and complicated set or agreements going back to a 1922 ruling, the Colorado River water is shared according to compacts, federal laws, court decisions and decrees, contracts, and regulatory guidelines collectively known as "The Law of the River."[7] Upper Basin states Colorado, New Mexico, Utah, and Wyoming and Lower Basin states Arizona, California, and Nevada are allocated different amounts of water. Mexico was added in 1944. Various sovereign nations like the Navajo have sued the federal government since 1999 arguing that tribal rights have been ignored. Changes to the apportioned amounts have been proposed as recently as 2008 as part of the presidential election campaign. The Law of the River is likely to change again as water becomes more valuable.

11.5 THE HETCH HETCHY NETWORK

San Francisco is well acquainted with disaster. The 8.3 magnitude earthquake and subsequent fires of 1906 are reminders to the city that disaster is always just around the corner. In more modern times, the city suffered heavy damage in the 1989 Loma Prieta earthquake. These natural

[6]https://en.wikipedia.org/wiki/Sfpuc

[7]https://en.wikipedia.org/wiki/Colorado_River_Compact

disasters have forced San Francisco to constantly hone the skills of its firefighters and emergency response personnel to deal with the unexpected:

> At 5:04 P.M., Tuesday, October 17, 1989, as over 62,000 fans filled Candlestick Park for the third game of the World Series and the San Francisco Bay Area commute moved into its heaviest flow, a Richter magnitude 7.1 earthquake struck. It was an emergency planner's worst-case scenario. The 20-second earthquake was centered about 60 miles south of San Francisco, and was felt as far away as San Diego and western Nevada. Scientists had predicted an earthquake would hit on this section of the San Andreas Fault and considered it one of the Bay Area's most dangerous stretches of the fault.
>
> Over 62 people died, a remarkably low number given the [rush hour] time and size of the earthquake. Most casualties were caused by the collapse of the Cypress Street section. At least 3,700 people were reported injured and over 12,000 were displaced. Over 18,000 homes were damaged and 963 were destroyed. Over 2,500 other buildings were damaged and 147 were destroyed.
>
> Damage and business interruption estimates reached as high as $10 billion, with direct damage estimated at $6.8 billion. $2 billion of that amount is for San Francisco alone and Santa Cruz officials estimated that damage to that county will top $1 billion.[8]

The water supply, however, was minimally impacted by the 1989 disaster. City workers sampled the quality of the water the next day and noted many breaks in lines, but no major disruptions in the availability of drinking water. The eight hills throughout the city lost power, and firefighters were forced to pump water from the bay to put out fires, but the city's 4-day supply of water remained intact:

> The greatest damage to the water system consisted of approximately 150 main breaks and service line leaks. Of the 102 main breaks, over 90 percent were in the Marina, Islais Creek and South of Market infirm areas. The significant loss of service occurred in the Marina area, where 67 main breaks and numerous service line leaks caused loss of pressure.[9]

The damage was minor when considering the size and complexity of the city's water system. Twelve gatemen run the whole system. "The system" contains over 1,300 miles of pipeline connecting 8,000 hydrants and 45,000 valves. It delivers 80 million gallonsday to 770,000 city dwellers. The SFPUC, which bills to 160,000 m, sells the surplus to another 1.6 million suburban users around the Bay Area.[10]

[8]The October 17, 1989 Loma Prieta Earthquake, http://www.sfmuseum.net/alm/quakes3.html#1989

[9]Memorandum to Tom Elzey, PUC General Manager from Art Jensen, Acting General Manager, November 21, 1989. Museum of the City of San Francisco.

[10]http://Sfwater.org

The major lesson learned from 1989 water supply damage was to buy more backup power systems. The earthquake tested the plumbing and purity of the water, not its availability. San Francisco was lucky because water kept flowing into the city from 175 miles away. The Hetch Hetchy valley and reservoir located in the Yosemite National Park supplies most of the city's water. What happens if this huge water resource dries up? This is the case of Hetch Hetchy, which experienced minor disruptions in 1997 and 2002.

11.5.1 Bottleneck Analysis

San Francisco maintains six municipal wells and 980 acres of lakes and land, but the bulk (65% or more) of its water comes from the pure lakes, reservoirs, and streams of the Hetch Hetchy. Hetch Hetchy is a network of 14 reservoirs, 22 pumping stations, several tunnels, and a number of treatment plants, filtration plants, and storage temples. This system delivers 400 million gallons of drinking water per day to 2.4 million customers. It is so big and complex that the first step is to identify the major components of the Hetch Hetchy network.

Figure 11.2 shows the expansive Hetch Hetchy network on top of a map of Northern California. This network consists of lakes, reservoirs, rivers, storage temples, treatment facilities, tunnels, and pipes needed to deliver water from mountain lakes to city dwellers.

Node and link robustness analysis suggests vulnerability due to an inadequate number of (redundant) links. Risk assessment should focus on this inadequacy. Because we are interested in the flow of water through these critical links, the nodes and links in Figure 11.2 are ranked by MBRA according to betweenness—the number of paths through each node/link. (The maximum number of paths turns out to be 402.)

Note there are two main links running horizontally across the Central Valley—the upper link is Hetch Hetchy's power transmission line, and the lower link is the pipeline and tunnel distribution link delivering water. Power is generated by several hydroelectric dams and then delivered to the Bay Area by a transmission line. Water is delivered through a pipeline system consisting of three pipes along some stretches on the way across the Central Valley.

From previous chapters we know that flow bottlenecks can be revealed through betweenness analysis. High betweenness indicates criticality because nodes and links with high betweenness support the most paths through the network. When the network is directed, as it is in this case, betweenness indicates the level of criticality of flow from source to sink.

There are seven sources of water from reservoirs and one storage temple—Hetch Hetchy, Lake Eleanor, San Antonio, Calaveras, Pilarcitos, Crystal Springs, and Sunol temple. Fragility comes from the limited number of linking pipes and twelve intermediate blocking nodes. Removal of any one of these blocking nodes segments the network into islands.

FIGURE 11.2 The Hetch Hetchy water and power supply network starts in the Hetch Hetchy region of Northern California and stretches 175 across California to the San Francisco Bay Area.

The highest-ranking betweenness values lie along the water pipeline passing through the Sunol treatment facility (#1), and continuing through the pipeline links around the southern end of the Bay, and then running north to San Francisco, via Silicon Valley. This series of high-betweenness pipes and nodes forms a *critical path* from source to destination.

This is a resilient network from the point of view of cascade failure, but is it resilient to flow disruptions? How critical are the critical paths? Accidental or human-caused hazards anywhere along these critical paths will have major consequences. According to the betweenness centrality metric, the most critical nodes and links from Figure 11.2 are:

1. Sunol treatment facility
2. Bay Division Junction
3. Junctions 2 and 3
4. Coast Range Tunnel
5. Pipes 1–2–3–4
6. San Joaquin Pipeline
7. New Don Pedro Reservoir
8. Palo Alto
9. Tuolumne River
10. Foothill Tunnel

All of these assets lie along a critical path from the collection source to the destination in San Francisco. Failure in any one of these nodes/links disrupts the flow of water. But flow resilience analysis seeks to determine if a CIKR network is able to tolerate outages in one part of the network by re-routing flow to alternative paths. Flow resilience depends on two properties of the directed network: tolerance for overloading links and availability of alternate links. Overloading occurs when the flow from a broken link is re-routed through a working path, and availability of alternate paths means there is more than one path from source to destination.

Figure 11.3 shows the results of flow resilience analysis. Due to its large number of alternative paths from source to destination, the Hetch Hetchy network is modestly resilient. When one link fails, it is rather easy to re-route the flows or to depend on alternative sources of water.

The graphs of Figure 11.3 plot flow PML risk versus overflow ratio on normal and log–log scales. Resilience is equal to the fractal dimension of the log–log scale plot. It is 1.08, here, which is moderate. But the top graph shows how resistant the network is to overloading. As overloading ratio increases along the x-axis, the PML risk drops, suggesting tolerance versus overloading ratio. Eventually, PML risk drops to zero as overloading exceeds 10.0—meaning a pipeline can tolerate a 10-fold overload without bursting. Flow resilience is the ability of a network to tolerate additional stress on its links due to overloading. As links fail, fewer alternative routes take on more overloading. And as they take on overloading, they become more likely to fail also. This cascading of overloaded

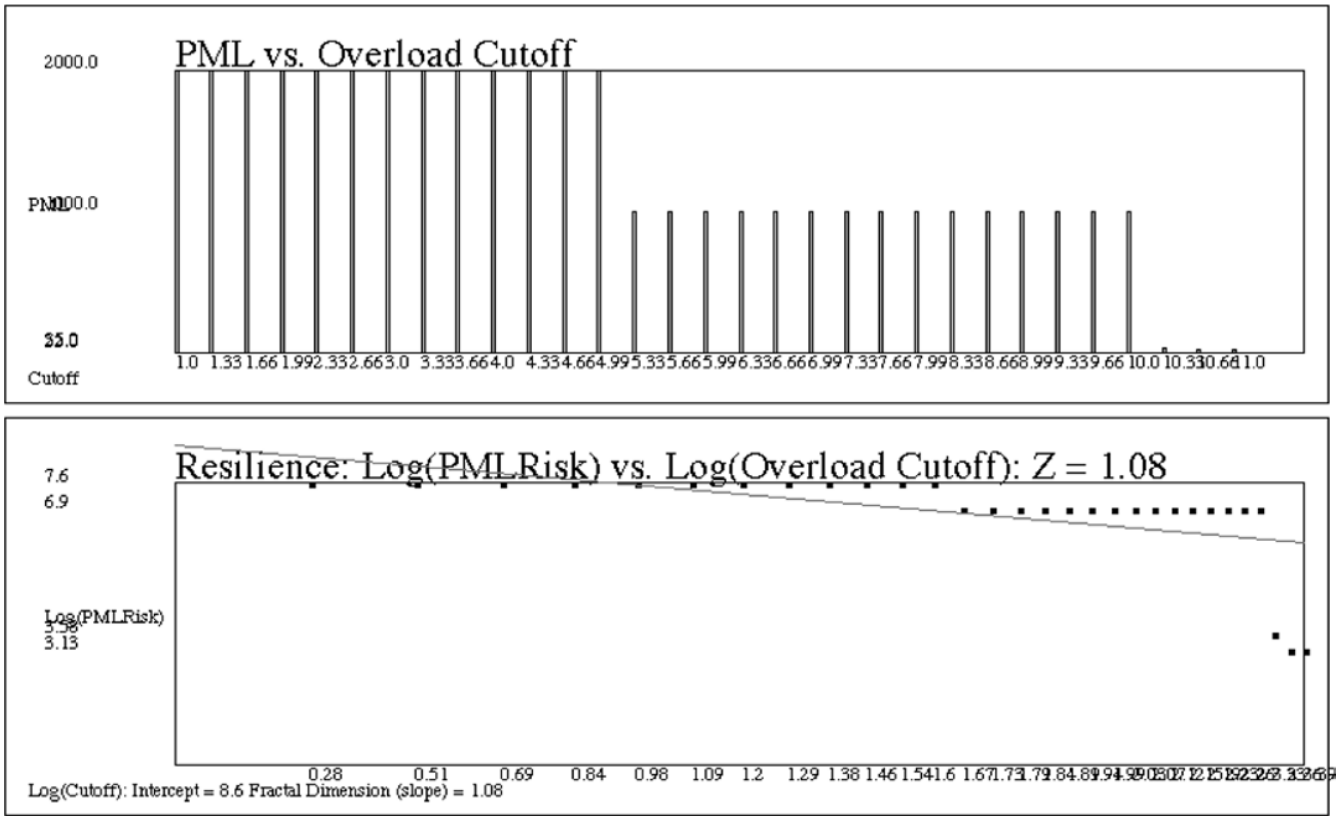


FIGURE 11.3 Flow resilience analysis shows the Hetch Hetchy network is modestly resilient to single link outages due to its tolerance for alternate routes.

links shows up in the top graph of Figure 11.7. Resilience is higher if the PML risk drops more rapidly with tolerance for overloading.

In fact, this network has a history of breakage. In November 2002 a leak in the critical path connecting Hetch Hetchy with the Bay Area treatment and distribution network cut the water supply to San Francisco in half. 210 to 240 million gallons/day stopped flowing to the Bay Area, but for only a few days. Fortunately, there is a 4-day supply of water "in the system," which buffered the effects of this accident. The network tolerated this break and continued to supply water to the city.

11.6 RISK ANALYSIS

The impact of a pipeline fault is likely to propagate downstream in the form of DoS—water ceases to flow. One might use downstream cascade simulation to estimate the downstream impact of a pipeline fault by assuming downstream assets fail because of upstream failures. Assuming a PRA risk model with TV equal to the probability of a subsequent failure downstream and C is the corresponding consequence in terms of loss of nodes, the simulation produces exceedence distributions as shown in Figure 11.4. Clearly, an investment

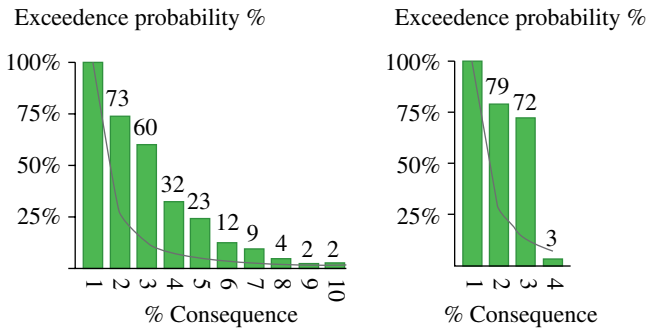


FIGURE 11.4 Simulation of downstream cascades caused by a random failure of a node or link in Figure 11.2 shows improvement after an investment in prevention (vulnerability reduction). Left: before prevention. Right: after vulnerability reduction.

in prevention reduces the long tail of the exceedence distribution, but it does not tell the whole story.

11.6.1 Multidimensional Analysis

Betweenness centrality identifies bottlenecks in the flow of water through the system. Connectivity identifies super-spreaders that magnify the spread of cascade failures. A combination of the two may provide a better

model of self-organization in the SFPUC water supply. Combining betweenness and connectivity centrality produces a ranking of nodes and links as follows:

1. Sunol treatment facility
2. Junctions 1 and 2
3. New Don Pedro Reservoir
4. Palo Alto
5. Bay Division Junction
6. Holm Powerhouse
7. Kirkwood Powerhouse
8. San Joaquin Pipeline
9. HH Power Junction
10. Pulgas temple

Alternatively, an upstream disruption of flow will have a downstream impact, so it may make sense to rank nodes and links according to betweenness and height (distance from the source). This metric yields a slightly different list:

1. Sunol treatment facility
2. New Don Pedro Reservoir
3. San Joaquin Pipeline
4. Foothill Tunnel
5. Bay Division Junction
6. Coast Range Tunnel
7. Holm Powerhouse
8. Kirkwood Powerhouse
9. Tuolumne River
10. Pipes 1–2–3–4

11.6.2 Blocking Nodes

Blocking nodes are the nodes that hold the CIKR network together. Removal of blocking nodes eliminates cascades, but it also eliminates continuity of flow. So in this case, blocking nodes are the minimum number of nodes required to keep the water and electricity flowing. Therefore, the blocking nodes of the SFPUC water and power network are highly critical.

Theory predicts 1/2.63 = 38% of the nodes are blocking nodes, but an automated brute-force method identifies 43%, or 12 nodes, as blocking nodes. Applying the automated algorithm for finding blocking nodes in the SFPUC water network yields the following critical nodes, in alphabetical order:

1. Cherry Tunnel
2. Coast Range Tunnel
3. Crystal Springs Tunnel
4. Don Pedro Reservoir
5. Foothill Tunnel
6. Holm Power
7. Lake Lloyd Reservoir
8. Junction 1
9. Junction 2
10. Power Substation*
11. San Andreas Reservoir
12. Sunol Valley Treatment

Limited resources may prohibit hardening of all of these nodes, but doing so would prevent cascade failures to both the water flow and power flow. The blocking node algorithm does not distinguish between pipelines and power lines, so the set of blocking nodes for water only excludes the power station node (indicated in the list above by an *).

11.7 HETCH HETCHY INVESTMENT STRATEGIES

In November 2002, San Francisco voters approved legislation to finance the largest renovation in the history of their water delivery system. The $3.6 billion capital program funded 77 projects to repair, replace, and seismically upgrade the water system's aging pipelines, tunnels, reservoirs, and dams. Did they spend the money wisely? Should the SFPUC invest more in prevention or response? Does the threat of terrorist attack change the strategy? These question are addressed by running a number of scenarios in MBRA:

- Risk reduction through prevention
- Risk reduction through response
- Risk and the rational attacker

The following results were obtained using hypothetical data for obvious reasons. This analysis is educational only and should not be construed as accurate or appropriate results for Hetch Hetchy. Some of the values were exaggerated to make an educational point.

Figure 11.5 summarizes the results of an MBRA network analysis versus investment in each of the scenarios above. Prevention is defined as retrofitting infrastructure against hazardous earthquakes and storms, adding fencing and CCTV cameras, and preventing ICS-SCADA exploits. Protection attempts to anticipate hazards before they happen and harden assets against potential damages. MBRA models the benefit of prevention by reducing vulnerability. (Remember risk is TVC.)

Figure 11.5 shows how risk declines as the prevention budget increases up to $1000 million. Risk declines versus investment because vulnerability declines. (Initially all vulnerabilities are set to 100%.) At about $100 million, the curve flattens out, suggesting a rapid diminishing return on

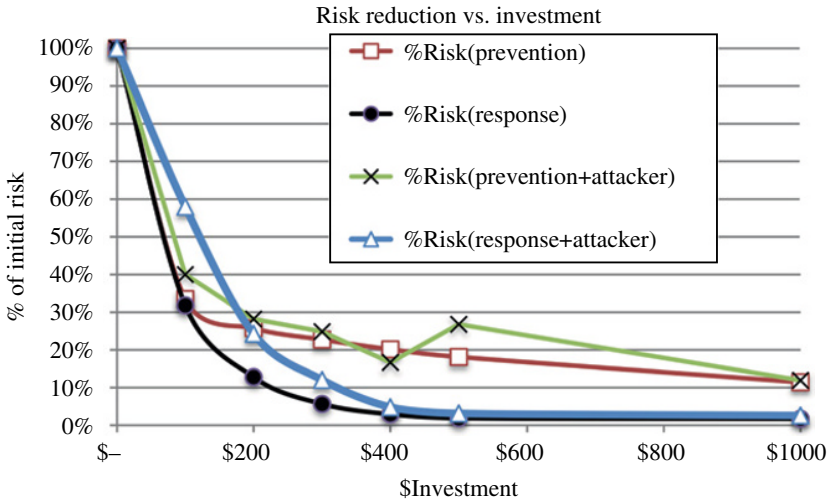


FIGURE 11.5 Risk versus investment in both vulnerability reduction (prevention) and consequence reduction (response) under different scenarios shows that investment in response is a slightly better strategy.

prevention investment. Prevention is less effective than response because of high prevention costs relative to consequences in the network model.

Investments in response typically involve investments in equipment and new technology to more rapidly respond to floods, fires, and other hazards. The goal of response funding is to reduce consequences. MBRA applies response funds to reduce consequences according to an exponential diminishing returns curve. Consequences range from tens of millions to $1000 million in the model, so consequence reduction can go a long way toward risk reduction. (Remember risk is TVC.)

Figure 11.5 shows a rapid decline in risk versus investment in response. In fact, it is the most effective strategy when investing more than $100 million. But the return quickly diminishes as investment approaches $400 million. Nonetheless, investment in response (consequence reduction) is the most effective strategy because consequence reduction is relatively inexpensive in the hypothetical data used in this illustration. See Table 11.3 containing the five most consequential assets in the Hetch Hetchy network model.

11.7.1 The Rational Actor Attacker

MBRA uses a *Stackelberg* optimization model to evaluate rational actor attacks on networks (see Appendix B for details). Stackelberg is a simple idea: the network defender allocates prevention and response resources to minimize overall risk, while a human attacker allocates attack resources to maximize overall risk. In terms of MBRA, a prevention budget is used to reduce vulnerability, a response budget is used to reduce consequences, and an attack budget is used to increase threat probability. MBRA iterates between risk minimization and risk maximization by shifting budgets from one asset to another. If equilibrium exists, a stalemate between minimization and maximization, then MBRA stops and recalculates T, V, and C using optimal allocations.

If equilibrium does not exist, MBRA outputs may oscillate back and forth, because more than one allocation satisfies the maximum–minimum requirement. For example, it is possible that many different patterns of allocation of prevention, response, and attack budgets produce an identical risk value. When this happens, MBRA outputs will change forever.

The Stackelberg algorithm assumes a rational actor—an attacker that is as rational as the defenders. Rational actors attempt to optimize expected utility—attacker maximizes risk and defender minimizes risk. But not all terrorists are rational actors. Some attackers are opportunistic more than rational. In fact, threat is often modeled as a combination of *intent* and *capability* rather than the probability of a rational action. Intent and capability are used in MSRAM, for example, to obtain the probability of threat–asset pairs.

Assuming equilibrium exists, two of the curves in Figure 11.5 show a higher risk versus investment when an attacker budget is applied at the same time as prevention and response budgets. Risk is generally higher because T is higher. Once again, risk is reduced more by response funding than prevention funding even when a rational actor threatens a network.

Figure 11.6 shows results of a rational actor attack on the Hetch Hetchy network. Risk is weighted by both betweenness and connectivity and summed over all nodes and links.

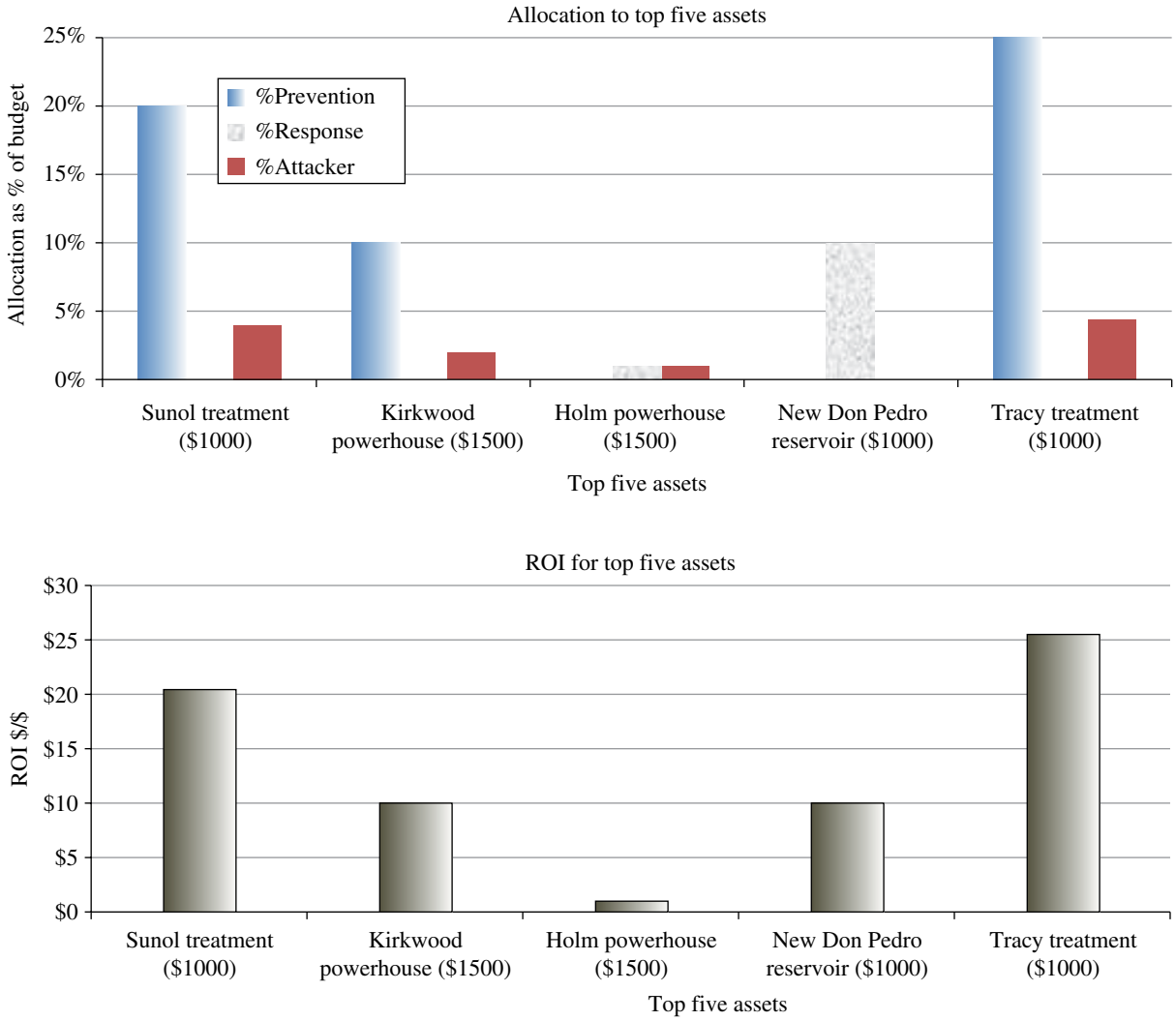


FIGURE 11.6 Results of Stackelberg optimization of defender and attacker allocations show that in both cases, players attempt to optimize return on investment (ROI) in order to optimize on risk.

Keep in mind that this is static risk, not dynamic or PML risk. The result of Stackelberg optimization is an assignment of threat T to each node and link such that static risk is as large as the attacker can make it and an assignment of vulnerability V and consequence C as low as the defender can make them to minimize static risk.

In this scenario, an attacker targets high-consequence assets because they increase risk more than low-consequence assets. Therefore, the attacker increases T for these high-value targets, and the defender attempts to reduce V and C. The process is iterative, first minimizing TVC by reducing VC, followed by maximizing TVC by increasing T. If an equilibrium point exists, iteration between attacker and defender allocations settles into a fixed point.

As a result of Stackelberg iteration, the top five assets are identified:

1. Sunol treatment facility (C = $1000 million)
2. Kirkwood Powerhouse (C =$1500 million)
3. Holm Powerhouse (C = $1500 million)
4. New Don Pedro Reservoir (C = $1000 million)
5. Tracey treatment facility (C = $1000 million)

Risk is optimal for both attacker and defender at $793 million (out of $19,450 million, initially), assuming prevention, response, and attack budgets are $100 million, $100 million, and $500 million, respectively. Allocation of budgets to prevention, response, and attack is shown in the

upper graph of Figure 11.6. Forty-five percent of the prevention budget is allocated to Sunol and Tracy treatment facilities. No prevention funds are allocated to Holm and New Don Pedro Reservoir. Why?

The lower graph in Figure 11.6 explains why allocations favor one asset over another. Funding goes to high ROI assets. Note that Sunol and Tracy treatment facilities return more risk reduction per investment dollar than the other assets. This is because MBRA allocates more prevention and response dollars to these high ROI assets. Further note that the attacker spends more on assets that do not leverage prevention dollars as well as response dollars. This is a consequence of Figure 11.5, which shows that prevention is less effective than response in terms of risk reduction. The rational attacker takes advantage of this weakness.

Both attacker and defender leverage ROI in two-party Stackelberg games such as this. Threat is increased when relatively large gains in risk are possible. Similarly, V and C are decreased when relatively large declines in risk are possible. The attacker tries to maximize T(VC) and the defender tries to minimize (T)VC, where the parentheses indicate holding x constant in (x) while varying T of VC. (ROI is determined by prevention and response costs and consequence.)

It is important that the risk analyst understand how a tool like MBRA arrives at its recommendations. Stackelberg optimization ignores all emotional attachments to CIKR and focuses narrowly on numerical values. It does not care if the CIKR is the Statue of Liberty or the Rock Island Bridge. It does not consider second-order consequences such as a drop in economic activity or loss of air travel revenues, unless the human analyst quantifies these secondary consequences in variable C. Most of all, MBRA ignores political and sociological issues that are difficult or impossible to quantify.

11.8 HETCH HETCHY THREAT ANALYSIS

The foregoing network analysis says the Hetch Hetchy water network contains critical nodes and links along critical paths from collection sources through a pipeline network with vulnerable tunnels and junctions eventually reaching treatment plants and the consumer. More specifically, the network analysis identifies high-betweenness nodes as critical, as well as highly connected nodes. The Stackelberg attacker–defender optimization further identifies upstream nodes such as the Holm and Kirkwood Powerhouses and key resources such as the New Don Pedro Reservoir as critical because of their high consequences. But network analysis does not include a threat–asset pair analysis. What threats should be considered and how should limited budgets be applied to reduce vulnerability to these threats? These questions are addressed here using fault trees.

When considering criticality due to betweenness, connectivity, and attacker risk, four assets rank high on all lists:

1. Sunol and Tracy treatment facilities
2. San Joaquin Pipeline and junctions such as Foothill and Coast Range Tunnels and Bay Division Junction
3. New Don Pedro Reservoir
4. Holm and Kirkwood Powerhouses

This list of critical nodes is narrowed down further to simplify the following threat analysis. As before, hypothetical values are used to protect the security of these real assets. Consider only the top four: Sunol treatment (Sun), San Joaquin Pipeline (SJP), New Don Pedro Reservoir (NDP), and Holm Powerhouse (Holm) as representative. These four assets lie on the critical path and represent the major asset types—treatment, pipeline, storage/collection, and power generation.

The following threat–asset pairs are analyzed as shown in Figure 11.7 using hypothetical values of T, V, and C and prevention/elimination costs listed in Table 11.4. These threats are representative only, and clearly they can be augmented by a much larger list.

- Sunol–SCADA, Sunol–CBRNE (chemical, biological, radiological, nuclear, and explosive weapons), Sunol–power outage
- San Joaquin Pipes–corrosion, San Joaquin Pipes–earthquake
- New Don Pedro–bomb, New Don Pedro–biological
- Holm Powerhouse–weather

In Table 11.4 all threats are assumed to be 50% and all vulnerabilities are assumed to be 100% initially. Pipeline damages due to earthquakes are assumed to be the most consequential in California, with bombs and weather events next in order of severity. The Holm Powerhouse could be damaged by flooding or extreme weather such as the massive flooding that occurred in 1861. Pipeline corrosion and power outages are assumed to be the least consequential of all even though they are responsible for frequent disruptions. Other threats that might be considered include forest fires and vandals in search of copper from power lines.

The most critical threat–asset pairs apply to treatment facilities such as the Sunol plant. See Chapter 10 for more on ICS-SCADA exploits. In addition to cyber exploits, the treatment facility is subject to closure due to a lack of power and also some kind of CBRNE attack. In fact, CBRNE attacks are not as unusual as one might expect.

11.8.1 Chem/Bio Threats

It is always difficult to estimate damages caused by an event that has yet to take place, but one common technique is to

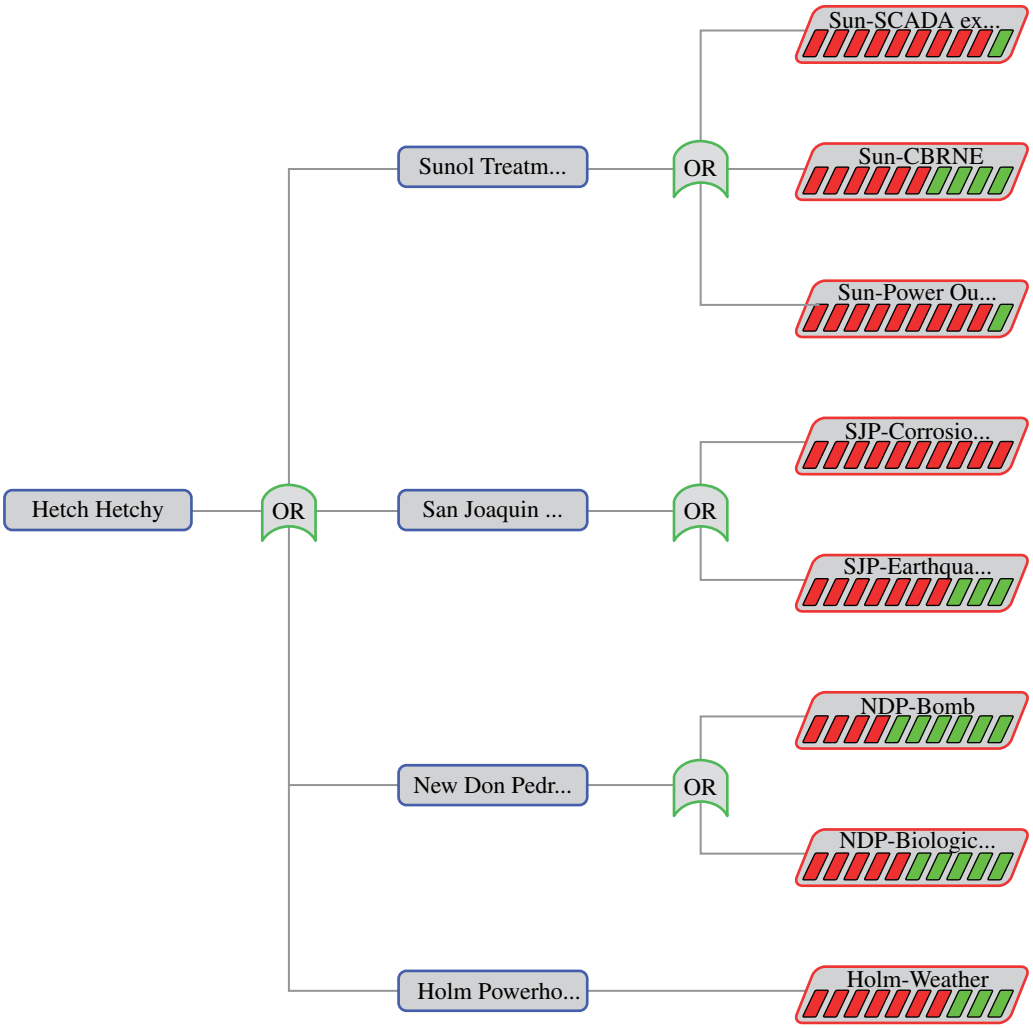


FIGURE 11.7 A fault tree model of Hetch Hetchy critical nodes identifies the most likely threat–asset pairs in the SFPUC water system.

TABLE 11.4 Hypothetical input values for the fault tree of Figure 11.7 indicates that earthquakes are the most consequential of all threats considered

Name	Threat (%)	Vulnerability (%)	Elimination cost $(millions)	Consequence $(millions)
SJP–earthquake	50.00	100.00	1000.00	2500.00
NDP–bomb	50.00	100.00	100.00	500.00
Holm–weather	50.00	100.00	40.00	100.00
Sun–CBRNE	50.00	100.00	15.00	50.00
NDP–biological	50.00	100.00	10.00	40.00
Sun–SCADA exploit	50.00	100.00	10.00	20.00
Sun–power outage	50.00	100.00	5.00	10.00
SJP–corrosion	50.00	100.00	10	10.00

look at similar incidents that have taken place in the past. For example, the largest ever chem/bio "attack" on drinking water occurred in Milwaukee, Wisconsin, in 1993. An unusually high volume of spring runoff was contaminated by cryptosporidium in fecal matter from cattle. The contaminated water entered the drinking water supply, which was not treated properly by the Milwaukee treatment plant. Cryptosporidium causes diarrhea in animals and humans.

An analysis of the 1993 Milwaukee cryptosporidium mishap suggests that chem/bio incidents are real, but perhaps not as devastating as we might think:

> Cryptosporidium came to national attention in 1993 in Milwaukee, Wisconsin, where 400,000 people were sickened. The protozoan was traced to a water filtration plant that served a portion of Milwaukee with drinking water. An investigation found there was a strong likelihood the organism passed through the filtration process and entered the water supply distribution system. The actual origin of this organism has been speculated to come from animal operations located in the tributaries of Milwaukee River. These tributaries drain directly into Lake Michigan, just north of where the water intake is located. [2]

Estimated consequences from the Milwaukee's cryptosporidium outbreak were $75–118 million [3]. While this was—and still is—the largest known biological incident to affect drinking water in the United States, the per-capita damages were modest, approximately $80/person for medical treatment and $160/person for loss of productivity. Rather conservative consequence estimates should be used when estimating the effects of a biological or chemical attack.

Chem/bio attacks on large bodies of water are not easy to do because of the diluting effect of lakes and reservoirs. It takes a large amount of contamination because the natural tendency of nature is to break down the molecular structure of chemicals and germs—which dilutes their effectiveness. In addition, the EPA has done a good job of regulating large water systems so that their treatment plants are equipped with chemical and biological detection and purification equipment. For these reasons, the consequences of a successful chem/bio attack on the Sunol treatment facility is relatively low compared with an earthquake, bomb, and weather damages. Similarly, elimination costs are typically low.

11.8.2 Earthquake Threats

Earthquakes are known to cause extreme damage to infrastructure in large cities. They are also known to do a lot of damage to pipelines. However, the consequences in terms of economic value are comparatively low. Nonetheless, California anticipates a 7.9 earthquake within the next 30 years, and much of the pipeline infrastructure has been put in place since the 1906 earthquake (8.2). In addition, pipes inside of tunnels may be impacted much more because of the tunnels themselves. Therefore, consequences and elimination costs are relatively high as shown in Table 11.4.

A nonprofit industrial organization called the Bay Area Economic Forum (BAEF) does studies to support the economic well-being of the San Francisco Bay Area. In October 2002, the BAEF released a report titled "Hetch Hetchy Water and the Bay Area Economy." This report estimated the impact of a 7.9 magnitude earthquake on the Bay Area, providing a sound basis for the estimated cost and damages expected of a major earthquake in this area of the country.

The BAEF report makes an impression: a 7.9 earthquake along the Hayward Fault would produce a loss in productivity and physical damage of $17 billion. Similarly, the combined economic and infrastructure damage caused by an earthquake along the San Andreas Fault would exceed $28 billion!

The area covered by Figure 11.7 is perhaps one-half of the area considered in the BAEF study. Additionally, buildings, highways, and pipelines have been earthquake hardened since the 2002 report. Therefore, an elimination cost of $1 billion was used in Table 11.4, because approximately one-half of the damage estimates used by the BAEF were attributed to economic losses, and the area residents have since spent $3.6 billion retrofitting infrastructure against earthquakes.

One notable exception may be the vulnerability of the pipeline passing through the Crystal Springs Tunnel near the Stanford Linear Accelerator off of Interstate 280. During the El Niño winter of 1996–1997, a landslide occurred on the northeast hillside above Polhemus Road in San Mateo County, which damaged homes and blocked Polhemus Road. The landslide temporarily buried the large water pipe running through the Crystal Springs Bypass Tunnel. The 96-inch pipeline transports an average of 90 million gallons of drinking water per day to communities in San Francisco and on the Peninsula, including San Mateo and parts of Silicon Valley. Ninety million gallons per day is 25% of the daily flow of Hetch Hetchy. The implication is that another storm, earthquake, or bomb attack could easily deny consumer access to 25% of the total water supply.

Another flaw in the model may be the absence of time-to-recover effects on consequence. The BAEF report estimates that repairing a pipeline takes 20 times as long as repairing a pumping station. Tunnels can take up to 30 times as much time. The fault tree model does not directly capture these delays, but delays can be quantified as economic loss. The fault tree lacks the expressive power to model time delays, but it does have the expressive power to model economic and productivity losses.

11.8.3 Allocation to Harden Threat–Asset Pairs

MBRA fault tree analysis applies the PRA equations, R = TVC, to each threat–asset pair shown in Figure 11.7 and then sums each threat–asset pair risk to obtain overall risk of $1615 million. Elimination costs are applied to each threat–asset pair to reduce vulnerability V according to an exponential diminishing returns curve. Therefore, risk also declines along a diminishing returns curve as shown in Figure 11.8. Of particular note is the mush slower decline in fault tree vulnerability—the probability that one or more threat occurs—compared to risk. Why is it more difficult to reduce vulnerability?

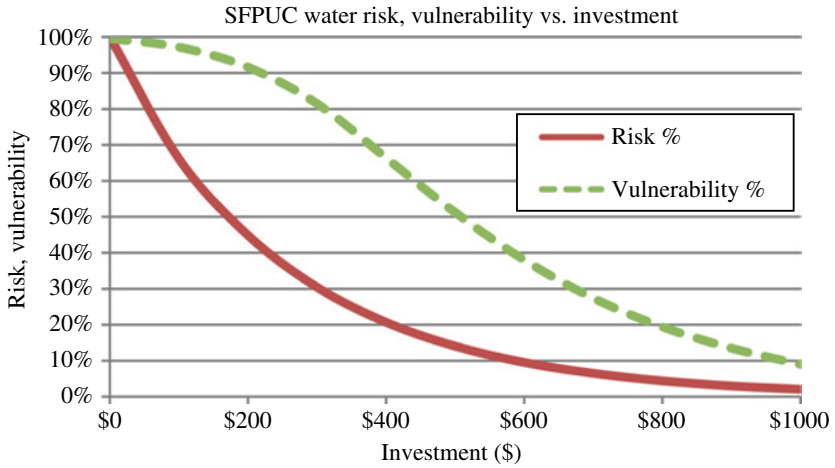


FIGURE 11.8 Risk reduction and vulnerability versus investment shows that risk declines much faster than vulnerability. An investment of $600 million is required to reduce vulnerability below 50%.

Investment in the Sunol–power outage pair is very inefficient compared with all other threat–asset pairs. For example, the ratio of initial risk to elimination cost is 0.50 for the Sun–power pair in Table 11.5 as compared with 1.25 for the SJP–earthquake pair. This means that investment in assuring a reliable power supply to the Sunol treatment facility returns very small reduction in vulnerability. For example, Sun–power vulnerability remains over 25% after a total investment of $600 million. (The fault tree uses OR-gate logic, which means that only one or more threats need occur to cause the entire fault tree to fail.)

The Hetch Hetchy system is only as secure as its least secure node or link. Therefore, a threat–asset pair such as the Sun–power outage pair that remains vulnerable after an investment means the entire system remains vulnerable. In this case, vulnerability is reduced from100 to 38% after $600 million is invested. The overall ROI is $2.44/$, which is a positive return.

Table 11.5 shows the results of an investment of $600 million to reduce V for the eight threat–asset pairs of Figure 11.7. Most of the money goes toward earthquake retrofitting (82%) with protection for a bomb threat against New Don Pedro Reservoir (11%). Therefore, 93% of the investment aims to harden the pipelines and reservoir. This strategy was produced by MBRA. Is it a good strategy? Human policy-makers might do well to modify the computer's results if the Sunol treatment facility is considered more important than MBRA says it is.

11.9 ANALYSIS

The Department of Homeland Security and the WaterISAC recommend the use of RAMCAP™ to perform risk analysis on pipelines, water systems, and industrial control systems. RAMCAP™ is an application of PRA: R = TVC, much like MBRA. But it does not perform resource allocation. Instead, the risk contribution of each threat–asset pair in an infrastructure is calculated, and then assets are ranked according to their risk. Risk ranking correlates well with MBRA resource allocation in this case, but in general, risk ranking is not guaranteed to yield an optimal allocation of limited funds:

> ***Risk ranking strategy****: Allocate risk elimination funds to the threat–asset pairs with the highest-ranking risk.*
> ***Risk minimization strategy****: Allocate risk elimination funds to the threat–asset pairs according to the highest ROI.*

Complex CIKR analysis shows that *self-organization* in the form of high betweenness makes the SFPUC water sector highly vulnerable to DoS attacks and natural disasters. A limited and hypothetical analysis of threats suggests that earthquakes are of major concern, but other assets such as tunnels may pose a greater risk because their destruction is easy and the time to repair them is high. A collapsed tunnel could lead to a long period of DoS. For example, the BAEF report estimates that repairing a Hetch Hetchy tunnel can take up to 30 times as much time as any of the other components.

Furthermore, the interdependencies among water, power, and transportation (airports) make water even more critical for the San Francisco Bay Area. It is conceivable that a normal accident that starts in the water sector could cascade to power and then to transportation. CIKR analysis of the Hetch Hetchy water and power network indicates a rather high resilience to cascade failure. Theoretically, vulnerability of individual nodes and links would have to exceed 25% to lead to a complex catastrophe. While this is highly unlikely, it is not impossible.

TABLE 11.5 Most of $600 million is allocated to harden pipelines against earthquake damage

Name	Threat (%)	Vulnerability (%)	Elimination cost $(millions)	Consequence $(millions)	Risk initial	Allocation $(millions)	Vulnerability reduced (%)	Risk reduced
SJP–earthquake	50.00	100.00	1000.00	2500.00	1250.00	492.19	10.37	129.58
NDP–bomb	50.00	100.00	100.00	500.00	250.00	64.27	5.18	12.96
Holm–weather	50.00	100.00	40.00	100.00	50.00	19.69	10.37	5.18
Sun–CBRNE	50.00	100.00	15.00	50.00	25.00	8.32	7.77	1.94
NDP–biological	50.00	100.00	10.00	40.00	20.00	5.94	6.48	1.30
Sun–SCADA exploit	50.00	100.00	10.00	20.00	10.00	4.44	12.96	1.30
SJP–corrosion	50.00	100.00	10	10.00	5.00	2.93	25.92	1.30
Sun–power outage	50.00	100.00	5.00	10.00	5.00	2.22	12.96	0.65

11.10 EXERCISES

1. Which of the following is DHS's mission in protecting the water sector?
   a. Environmental impact on water supplies
   b. Drinking water supplies
   c. Agricultural water supplies
   d. Industrial water supplies
   e. All of the above
2. When did responsibility for water security transfer from the Public Health Service (PHS) to the US EPA?
   a. 1914
   b. 1962
   c. 1974
   d. 2002
   e. 2003
3. Why did regulation of water move from PHS to EPA?
   a. Emphasis shifted from biological to environmental contamination.
   b. Emphasis shifted from biological to terrorism.
   c. The EPA had more money.
   d. PHS was abolished.
   e. Emphasis shifted from chemical to environmental contamination.
4. The Bioterrorism Act of 2002 extends the SDWA of 1974 as follows:
   a. Includes acts of terrorism
   b. Requires vulnerability assessments
   c. Establishes the WaterISAC
   d. Specifies prison term penalties
   e. All of the above
5. Which of the following are critical nodes in the SFPUC water system (Hetch Hetchy) as determined by network analysis?
   a. New Don Pedro Pipeline
   b. The ICS-SCADA system
   c. Hetch Hetchy and Lake Lloyd Reservoirs
   d. Sunol treatment facility
   e. Merge #2 and Merge #3
6. Optimal resource allocation finds the best allocation of budgets by maximizing:
   a. Threat
   b. Vulnerability
   c. Risk
   d. ROI
   e. Consequence
7. Resource allocation by risk ranking guarantees the following:
   a. Risk minimization
   b. Optimal allocation of resources
   c. Maximum ROI
   d. Threat minimization
   e. None of the above
8. The foundation of the water sector's safety and security is:
   a. The Bioterrorism Act of 2002
   b. The 1974 SDWA
   c. PPD-63
   d. PPD-21
   e. The Homeland Security Act of 2002
9. Stackelberg game theory finds the best attacker and defender allocation by:
   a. Predicting future attacks
   b. Maximizing threat and minimizing vulnerability
   c. Minimizing threat and maximizing vulnerability
   d. Maximizing threat, vulnerability, and consequence
   e. Minimizing threat, vulnerability, and consequence
10. Earthquake experience has shown that the water supply is most vulnerable to:
    a. Broken pipes
    b. Contamination of the lakes and reservoirs
    c. Collateral fires and explosions
    d. Collapsing tunnels
    e. Collapsing freeways
11. The main lesson learned from San Francisco earthquakes is:
    a. Pipes are vulnerable to earthquakes.
    b. Drinking water is no longer potable.
    c. Collapsing tunnels block the flow of water.
    d. 80% of the water is used for agriculture.
    e. Backup power is essential.
12. The largest water supply contamination disaster in the United States was:
    a. Hetch Hetchy, November 2002
    b. Milwaukee, Wisconsin, cryptosporidium contamination in 1993

c. The Loma Prieta earthquake in the San Francisco Bay Area
d. Hurricane Fran in 1996
e. Hurricane Dennis in 1999

13. In terms of time delays caused by the time to repair a water sector component, the Hetch Hetchy water supply is most vulnerable to:
a. Broken pipes
b. Collapsing tunnels
c. Collapsing freeways
d. Broken pumps and gates
e. Insufficient budget

14. Which of the following is the least interdependent with the water sector?
a. Transportation
b. Power
c. Agriculture
d. Silicon Valley industry
e. Public health

15. The major lesson learned from the 1989 earthquake in the San Francisco area relative to the water supply:
a. Buy more backup power systems
b. Retrofit power transmission lines
c. Backup the water supply to the airport
d. Harden tunnels
e. Duplicate treatment facilities

11.11 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. Water safety and security has gone through several stages from concern for purity to concern for terrorism. Does the rise of malware attacks against ICS add to this evolution? Why or why not?

B. Explain how global climate change is likely to impact water security. Does climate change add to the evolution from concern for purity to concern for terrorism and malware? Explain your answer.

C. If you were going to make Hetch Hetchy more resilient, what would you do, and why?

D. Does cascade resilience make sense when applied to a flow network such as Hetch Hetchy? Why or why not?

E. The average connectivity of a node in the SFPUC water supply network is 2.21 and its spectral radius is 2.63. Does this mean it is highly self-organized? Explain why or why not?

SIDEBAR 11.1 HISTORICAL TIMELINE FOR THE EVOLUTION OF WATER SAFETY AND PREVENTION OF TERRORIST ATTACKS ON DRINKING WATER

1880s—Louis Pasteur develops germ theory and notes that water is a vector

1885—Dr. John Snow proves cholera transmitted by drinking water

1914—US Public Health Service (PHS) sets standards for the *bacteriological* quality of drinking water

1925, 1946, and 1962—US PHS revises standards. The 1962 revision called for the regulation of 28 substances and established the most rigorous standards until 1974

1960—US PHS study shows that only 60% of drinking water met PHS standards

1972—US PHS study of Mississippi River reveals 36 *chemicals* contaminating drinking water processed by treatment plants

1974—Safe Drinking Water Act (SDWA) establishes foundation of modern regulations for protecting the purity of water and water systems. Enforcement transferred to US EPA

1986, 1996—Revisions to SDWA of 1974.

1993—Cryptosporidium outbreak in Wisconsin kills over 50 people and infects 400,000 consumers of public water

May 22, 1998—President Clinton signed Presidential Decision Directive 63 identifying drinking water as one of America's critical infrastructures

June 12, 2002—President Bush signs into law the Public Health Security and Bioterrorism Preparedness and Response Act of 2002

August 1, 2002—US EPA completes the *classified* Baseline Threat Report describing likely modes of terrorist attack and outlining the parameters for vulnerability assessments by *community* water systems

December 2002—Water Information Sharing and Analysis Center (WaterISAC) becomes operational.

March 31, 2003—Water systems serving more than 100,000 people submit vulnerability assessments to the US EPA

December 31, 2003—Water systems serving between 50,000 and 100,000 people are required to submit vulnerability assessments to the US EPA

June 30, 2004—Water systems serving between 3,300 and 50,000 people are required to submit vulnerability assessments to the US EPA

2014–2016—Water Infrastructure Finance and Innovation Act (WIFIA)

REFERENCES

[1] U.S. Environmental Protection Agency. *25 Years of the Safe Drinking Water Act: History and Trends*. Report No. US-EPA 816-R-99-007, December 1999. Available at http://www.epa.gov/safewater/consumer/trendrpt.pdf. Accessed June 29, 2014.

[2] Holman, R. E. *Cryptosporidium: A Drinking Water Supply Problem*. Water Resources Research Institute of the University of North Carolina. Special Report No. 12, November 1993.

[3] Corso, P. S., Kramer, M. H., Blair, K. A., Addiss, D. G., Davis, J. P., and Haddix, A. C. Cost of Illness in the 1993 Waterborne Cryptosporidium Outbreak, Milwaukee, Wisconsin, *Emerging Infectious Diseases*, 9, 4, April 2003, pp. 426–431.

12

ENERGY

Energy—derived from fossil fuels, wind, solar, and nuclear—propels everything. It runs power plants, heats our homes, powers our cars, and air-conditions our offices. Without energy, the telephone and the Internet would not work, and the modern conveniences we take for granted vanish. The United States consumes roughly 100 quads of energy/year. This is equivalent to 17.5 billion barrels or 735 billion gallons of crude oil. Approximately, 20% of this energy comes from coal, 23% from natural gas (NG), and 35% from petroleum. Figure 12.1 further indicates that 38% is converted into electricity to run computers, homes, offices, and factories. Transportation in the form of cars, trucks, busses, airplanes, ships, and trains consumes 27%. Incredibly, 55% is lost due to conversion, transmission, and nonconservation inefficiencies. For example, an incandescent light bulb based on Edison's invention wastes 95% of the energy needed to light up a room.

The energy sector is transitioning to renewable fuels such as wind, solar, tidal, and geothermal. This transition will take 40–50 years, but radically alter another infrastructure—transportation. Automobiles, trucks, trains, and airplanes are likely to be running on electrons, hydrogen, or some fuel other than gasoline by 2030. Conversion of the energy sector from fossil fuels to some alternative such as solar will most likely bring major sociopolitical and economic shifts with it.

This chapter examines critical nodes of the fossil fuel supply chains[1] that deliver much of the energy consumed by the United States. We illustrate how this supply chain works through case studies: the Powder River coal supply chain, the Gulf of Mexico to Northeastern US supply chain, and the Northeast storage facility located in New Jersey. The following summarizes the result of these analyses:

- *Energy versus power*: Energy is the ability to do work; power is the rate of doing work. Therefore, power is measured in units like kilowatts (kW) and horsepower (hp) (ft-lb/s), and energy is measured in units like kilowatt-hour (kWh). The mathematical relationship is $energy = power \times time$. The energy sector is concerned with extraction and delivery of fuels to power plants, and the power sector is more concerned with producing power from power plants and delivering it to consumers.
- *Coal supply chain*: This critical supply chain is highly dependent on transportation in the form of rail delivery of coal to power plants. The United States leads the world in

[1]According to Dr. Warren H. Hausman of Stanford University, "The term supply chain refers to the entire network of companies that work together to design, produce, deliver, and service products." http://www.supplychainonline.com/cgi-local/preview/SCM101/1.html

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

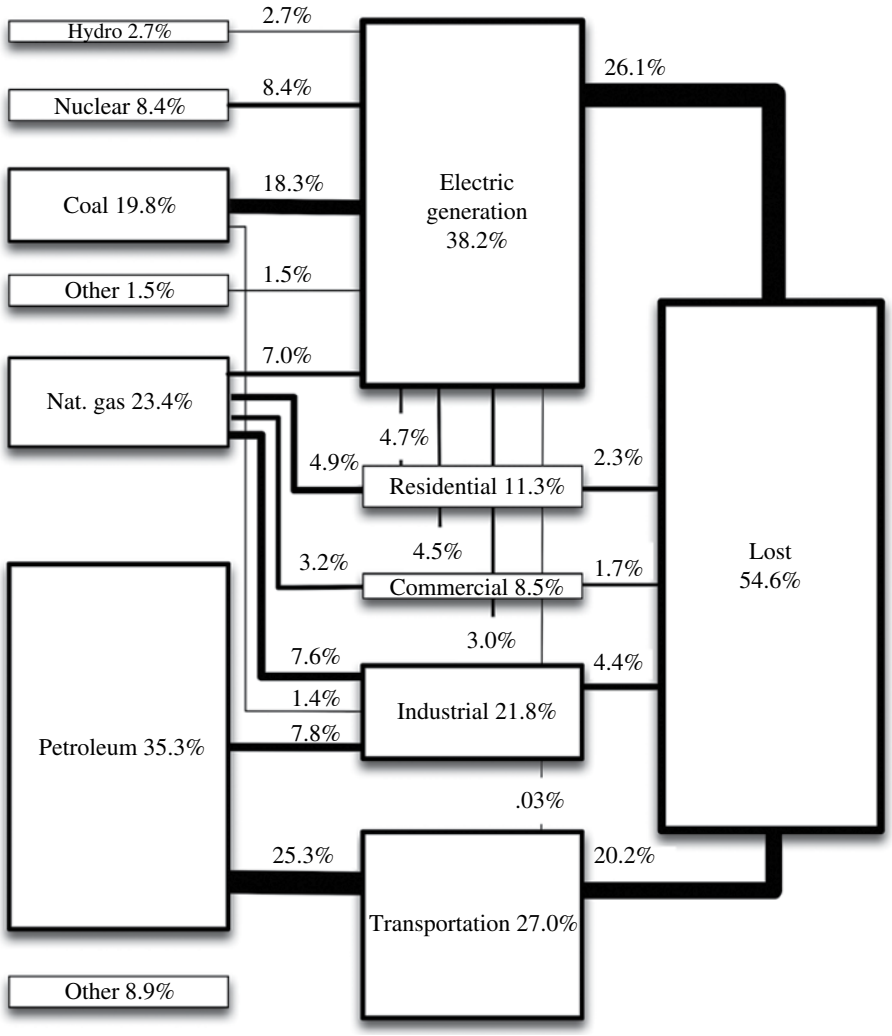


FIGURE 12.1 Most energy consumed in the US United States comes from coal, natural gas, and petroleum—fossil fuels—and over half is lost due to inefficiency.

coal reserves. The Powder River Basin is the largest source of coal in the United States, and its delivery depends on rail service to power plants. While the United States is highly dependent on coal as a fuel, environmental regulations continue to reduce the availability of coal for generation of electrical power. Coal is a declining source of energy for the United States, but not China.

- *Gas and oil supply chains*: Most US energy comes from vast oil and NG supply chain networks that are highly self-organized—they have high betweenness centrality and low levels of robustness. They are vulnerable to disruptions of their large and unique transmission pipelines that form supply chains from Canada, the Gulf of Mexico, and foreign sources.
- *Critical assets*: Major components of oil and NG supply chains are wellheads, refineries, transmission pipelines, storage, distribution pipelines, and SCADA. Supplies are vulnerable to disruption because of highly clustered refineries, pipeline ruptures, and storage facility damage.
- *Competitive exclusion*: Ownership of gas and oil refineries and transmission pipelines is highly concentrated. Dense clusters of refineries, limited transmission links, and concentration of storage terminals characterize energy supply chain networks. Robustness is limited because few assets are redundant.
- *Vulnerabilities*: Coal is vulnerable to transportation disruptions, and oil and NG supply chains are vulnerable to disruption because of asset clustering and low robustness of transmission. They are also heavily dependent on power (for running pumps and SCADA).

12.1 ENERGY FUNDAMENTALS

Most people are ignorant of perhaps the most important CIKR sector to modern civilization—the energy sector. Energy has become a commodity like water and the air we breathe, but like water and air, modern civilization would come to a halt in less than a month without a continuous supply of energy. A sudden cessation of gasoline, coal, or electrical power, for example, would throw the world back 500 years to medieval times. And yet, few consumers understand how this sector works. Even fewer understand how energy is measured, produced, delivered, and consumed. Taking this CIKR for granted may be a major mistake, however.

As a simple illustration, consider the architecture of the US energy sector as shown in Figure 12.1. The United States consumes approximately 18 million barrels of oil/day (2012), representing about 27% of all energy consumed. To gauge how large this number is, consider the energy content of one barrel of oil, which contains 42 gallons of crude oil. (This crude is converted into 44 gallons of gasoline, which is roughly equivalent to 1 week of fuel for a typical consumer.) The typical US consumer uses approximately 54 barrels of energy/year.

Energy is not created or destroyed, except by nuclear reaction. Instead, it is converted. A coal-burning power plant releases energy by burning coal and converting it into heat and motion by heating water to make steam, which drives a rotating turbine. The kinetic energy of the turbine is further converted into moving electrons by a generator. Thus, the energy stored in coal is converted into other forms of energy—heat, kinetic, and electrical—so it can be transmitted to where it is converted back into movement of an electric car, lighting of a house or office, or computation inside of a computer or cell phone.

Energy is not created or destroyed, but some of it is wasted in conversion. For example, burning coal is roughly 30% efficient, meaning that 70% of coal's pent-up energy is lost during conversion to heat, light, and chemical reaction. Sadly, most energy is lost due to conversion along the energy supply chain that extends from the coal mine to the home or office.

Energy is measured in terms of work performed over some time period, so energy and work are equivalent. On the other hand, power is work performed per unit of time. It is the rate of doing work. The difference is subtle so here is an example. Most everyone in the West is familiar with horsepower. James Watt defined horsepower as the amount of work a horse could do over a certain amount of time. He concluded that one horse could lift 33,000 lb of coal out of a mind in a minute. For example, if Watt's horse lifted 330 lb of coal 100 ft in a min, or 1,000 lb 33 ft in 1 min, his horse exerted 1 hp, regardless, because 330(100) = 33(1,000) = 33,000 ft-lb/min.

The relationship between power and energy (work) is

$$\text{Energy} = \text{power} \times \text{time}$$

Therefore, the amount of work done by Watt's horse depends on time: if his horse exerts 33,000 ft-lb/min for an hour, 33,000(60 min/h) = 1,980,000 ft-lb of work is done. If the horse works for 2 h, 3,960,000 ft-lb of energy is transferred from the horse to the coal (the coal has higher potential energy as a result of being lifted out of the mine).

There are many different measures of energy and power. In the energy sector, most technologists prefer to measure *energy* in kWh in place of foot-pounds (ft-lb), and *power* in kW instead of hp. One kW equals 1.34 hp, for example. A 100 W light bulb equals 0.134 hp, an electric car with 100 kW motor equals 134 hp, and so on.

The amount of work made possible by a fuel depends on its *energy density*. Table 12.1 lists a few popular fuels and their energy densities for comparison. Electrical energy is typically measured in kWh—a 100 W incandescent light bulb burning for 10 h converts 1 kWh of energy into heat and light. Modern US urban dwellers convert 500–1000 kWh of electricity into heat, cooling, computation, and communication, for example, per household, per month.

Mechanical energy is typically measured in British thermal units (BTU) instead of kWh. 1 kWh is approximately 3412 BTU. An electric car with a 100 kWh battery stores 341,200 BTU of energy when fully charged. If its efficiency is 3 kWh/min, the electric car can travel 3(100) = 300 miles on a charge. How long would it take? It depends on the power of its electric motor. How big is the battery? It depends on the energy density of the chemicals used to store electrons in the battery.

The United States converts approximately 100 quads/year of energy from all forms of fuel (see Fig. 12.1). A quad is an extremely large number equal to 10^{15} Btu, 8,007,000,000 gallons of gasoline, or 293,083,000,000 kWh. The United States uses nearly 100 times this amount of energy. The entire planet converted fuel of one kind or another into 446 quads in 2004; therefore, the United States used (95/446) = 21% of all energy captured by the planet in 2004.

TABLE 12.1 Energy density of familiar fuels: Uranium-235 is off the scale, while lithium-ion battery storage barely registers. Fossil fuels are comparatively dense sources of energy, but also produce large volumes of greenhouse gases such as CO_2

Fuel	Density (kWh/Gal)	CO_2 emission (lb/million BTU)
Li-ion battery	3	0
Natural gas	27	117
Gasoline	38	157
Coal	76	216
U-235	1,500,000,000	0

Over half of all energy produced in the United States is wasted—largely due to inefficiency. Electrical power transmission is the most wasteful followed by gasoline consumption. Together, they account for 46% of all energy—losses that escape in the form of heat, friction, and so on. This fact alone may drive future policy decisions regarding the CIKR sector. For example, conversion of personal transportation from the internal combustion engine (ICE) to an alternative such as fuel cells or electric motors can reduce the 20% loss to perhaps 5%, because electric motors are 3–4 times more efficient than ICE. As another example, consider the possibility of distributed generation, whereby electric power generation is brought closer to its point of consumption, thereby reducing transmission losses. Solar panels on residential homes and shopping malls, for example, reduce the need for long-haul transmission lines—the major source of electrical energy loss.

These and other transitions brought on by new technologies and new energy policies will radically alter the architecture of the energy sector. But consumption is unlikely to diminish because energy is the engine of economic prosperity and security. The confluence of increasing demand and diminishing inefficiencies levels the slope of future energy needs. Figure 12.2 summarizes where we have been and where we are going with respect to this CIKR. Projected energy demand is most likely to track population and economic trends, which are typically 2.5–3.5% per year.

For the past 100 years, civilization has not only increased its demand for energy but also increased the rate of increase. Energy demand has accelerated because of greater mobility (transportation), adoption of technology (computers and cell phones), and increased population (from 1.5 to 7 billion in the past century). The supply of energy will hit an inflection point sometime in the twenty-first century, for a number of reasons. First and foremost, fossil fuels are becoming scarce and more expensive to exploit. Second, technology advances are likely to provide economically and environmentally more desirable alternatives such as solar and fuels derived from plants and new processes.

The following historical and contemporary analysis is unlikely to persist beyond the decade because population, technology, and economics will force a radical revamping of policy and strategy in the United States and abroad. Security and safety considerations are also likely to undergo radical modification as the energy sector transforms along one of several possible development directions outlined below.

12.2 REGULATORY STRUCTURE OF THE ENERGY SECTOR

Congress delegates oversight of gas, oil, and electric power to the Federal Energy Regulatory Commission (FERC). FERC must work with the EPA, the Department of Energy, and the Office of Pipeline Safety (OPS) positioned within the Pipeline and Hazardous Materials Safety Administration (PMHSA), which in turn is positioned within the Department of Transportation (DOT). The Atomic Energy Commission (AEC) is responsible for oversight of nuclear power plants.

12.2.1 Evolution of Energy Regulation

FERC evolved out of Federal Power Commission (FPC) established by Congress in 1920 to regulate hydroelectric

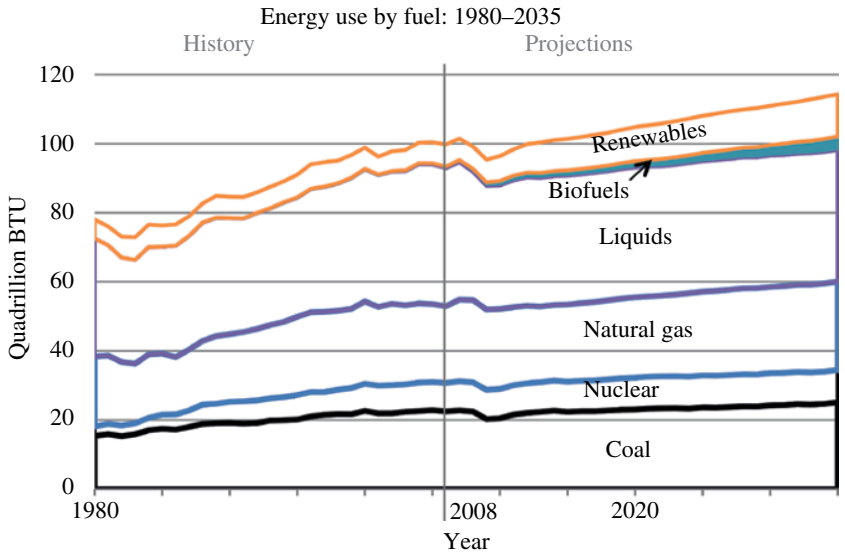


FIGURE 12.2 Past, present, and future energy consumption in the United States: fossil fuels will remain the most consumed energy for the foreseeable future. Source: From Ref. [1].

projects. But FPC quickly evolved as legislation increased federal control. A brief history as it pertains to the gas and oil industries is summarized below[2]:

- *Regulation of sale and transportation*: The Federal Power Act of 1935 and the Natural Gas Act (NGA) of 1938.
- *Regulation of NG facilities*: 1940 amendments to the NGA and 1954 Phillips Petroleum v. Wisconsin
- *Interstate commerce*: In 1967, intrastate utilities became jurisdictional if they connected their supply lines to others outside of the state.
- *FPC becomes FERC*: Congress reorganizes and expands FPC in 1977 as FERC.
- *Unified interstate commerce*: The 1978 National Energy Act (NEA) unifies intra- and interstate gas markets.
- *Deregulation*: 1985 FERC Order 436 and 1992 FERC Order 636 opens pipeline to competitors and introduces price controls.

The energy sector has been shaped by a century of regulation and deregulation. It forms an industrial commons balanced between the forces of Gause's law and competitive exclusion and the tragedy of the commons. Left to free-market forces, this commons would evolve into a monoculture and monopoly. Left to an over-regulated command economy, this commons would die from the ravages of the tragedy of the commons. In its early years, the federal government treated this sector like a natural monopoly. Since Congress changed its stance on regulation through the Public Utility Regulatory Policies Act (PURPA) of 1978 and Energy Policy Act (EPACT) of 1992, the energy commons evolved somewhere in between. Pipelines are open to competitors, but there is a relatively high concentration of pipeline ownership and operation. Energy costs are low for consumers, but government regulates (and limits) profitability of the industry.

12.2.2 Other Regulations

PMHSA regulates the safe, reliable, and environmentally sound operation of the nation's 2.6 million miles of gas and oil pipeline and the nearly 1 million daily shipments of hazardous materials by land, sea, and air. It consists of two offices: the OPS and the Office of Hazardous Materials Safety (OHMS).

Gas and oil pipeline companies are considered *common carriers*, which means they must operate as all interstate commerce companies: they must provide nondiscriminatory access to their networks. But the energy supply chain overlaps other domains because of its impact on the environment and dependence on transportation. Operational dependencies lead to complex regulatory controls. The NG and oil supply chains cross many regulatory boundaries as well as geographical boundaries.

The Natural Gas Pipeline Safety Act (NGPSA) of 1968 authorized DOT to regulate pipeline transportation of various gases, including NG and liquefied natural gas (LNG). As a consequence, DOT created OPS, but the emphasis was on safety, not regulation. The National Environmental Policy Act (NEPA) of 1969 made the FPC responsible for reporting environmental impacts associated with the construction of interstate NG facilities. And then the task of coordinating federal efforts to cope with electricity shortages was taken from the FPC and given to the Office of Emergency Preparedness in 1970. To complicate matters even more, the FPC was converted into FERC in 1977, and the NEA of 1978 that includes the Public Utility Regulatory Policies Act (PURPA) required gradual deregulation of NG. In 1979 Congress passed the Hazardous Liquid Pipeline Safety Act (HLPSA), which authorized the DOT to regulate pipeline transportation of hazardous liquids.

The EPAC of 1992 required FERC to foster competition in the wholesale energy markets through open access to transmission facilities. By 1996 FERC issued a series of orders forcing common carrier companies to carry electricity, NG, or petroleum products from a variety of competing suppliers. This re-regulation of the energy sector was euphemistically called "deregulation."

By 2003 the regulatory structure was cloudy at best. FERC handled regulation, DOT/OPS handled pipeline safety, and EPA handled hazardous materials. The DOE provided information on the energy sector through its EIA (Energy Information Agency at www.eia.doe.gov). HSPD-7, issued by President Bush in December 2002, distributes responsibility for the energy sector across DOE, DHS, DOT, and EPA. PMHSA was created under the Mineta Research and Special Programs Improvement Act of 2004, combining OPS and responsibility for hazardous material spills.

FERC sets prices and practices of interstate pipeline companies and guarantees equal access to pipes by shippers. It also establishes "reasonable rates" for transportation via pipelines. These charges typically add a penny or two to the price of a gallon of gasoline, for example. But FERC is not responsible for safety, oil spills, or the construction of the energy supply chain. Safety is regulated by PMHSA, except for nuclear energy, which falls on DOE and the AEC.

For example, when a pipeline ruptured and spilled 250,000 gallons of gasoline into a creek in Bellingham, Washington, in June 1999, OPS stepped in. This accident killed three people and injured eight. Several buildings were damaged, and the banks of the creek were destroyed along a 1.5 mile section.

Is there any difference between NG and oil when it comes to regulation? FERC is empowered to grant permission for anyone to construct and operate interstate pipelines, interstate storage facilities for NG, or LNG plants. It also handles requests by anyone who wants to abandon facilities when, for example, pipelines get old and need to be upgraded or

[2]http://www.ferc.gov/students/ferc/history.asp

replaced with a new pipeline. But FERC does not regulate local distribution pipeline companies—state public utility commissioners regulate them.

12.2.3 The Energy ISAC

The mission of the Energy ISAC (E-ISAC) is to provide threat and warning information to member companies in the energy sector. The federal government funds it, but a third-party company operates it from an undisclosed location. Its members are the companies that produce NG, oil, coal, and so on. The E-ISAC is not the same as the ES-ISAC (Electricity Sector ISAC).

According to www.energyisac.com, the E-ISAC provides its members with:

- Information on threats and vulnerabilities (physical, cybersecurity, and interdependencies).
- How to respond to threats (both physical and information security tips).
- A forum for members to communicate best practices in a secure environment.

12.3 INTERDEPENDENT COAL

Coal has been a cheap and plentiful source of energy for thousands of years and will continue to be a major source throughout the world. Its use in the modern age accelerated in the 1880s in concert with the rise of the industrial revolution. Coal powered the rapid rise of steel, rail, transportation, skyscraper, petrochemical, and electric power industries. Its relatively high energy density makes it an attractive source of energy (see Table 12.1). Without coal, the modern age would not be very modern.

Global reserves of oil will last about 50 years at current consumption rates. Coal, on the other hand, will last about 112 years at current consumption rates.[3] The United States has the largest reserves of coal (22%), and China has 12%. The United States produced 14% of the global supply in 2011 and consumed 12%. But China produced and consumed 50% of all coal produced in 2011. China is currently the largest market for coal.

Coal produces 20% of all power consumed in the United States, but this ratio is changing because of the high CO_2 emissions caused by burning coal (see Table 12.1). Coal-powered power plants produced 30% of output from all power plants. An EPA proposal in 2013 cuts CO_2 emissions from power plants by nearly 40% (from 1800 to 1100 lb/MWh). This regulation poses the largest threat to coal as a source of energy.

[3]http://www.worldcoal.org/coal/where-is-coal-found/

12.3.1 Interdependency with Transportation

Surprisingly, the largest source of coal in the United States is Wyoming, not West Virginia. Two mines—the Black Thunder and North Antelope Rochelle mines in Wyoming—produce almost as much coal as West Virginia. Over half of the coal produced in the United States is produced in the Western coal region, and Wyoming is the largest producer. Nine of the top 10 US mines are located in Wyoming.[4]

One region in Wyoming is particularly critical to the coal supply chain. Approximately 68% of US coal comes from the Powder River Basin region. But markets for this coal are spread all over the globe. How does Powder River Basin coal reach these markets? In general, 72% of mined coal reaches the power plants that consume it by rail, 11% by barge, 10% by truck, and 9% by pipeline. Powder River Basin coal depends on rail, because only rail can move such large volumes of coal, economically.

The Powder River Basin coal supply depends on Union Pacific (UP) and Burlington Northern Santa Fe (BNSF) to move over 300 million tons of coal to power plants in the Midwest and westward to markets in Asia (see Fig. 12.3).[5] These trains are dedicated to the task—they haul 100–125 cars of Powder River Basin coal to electric utilities and then return back to the mines empty. BNSF, for example, supplies 10% of all electrical power by transporting coal to power plants.

A 103 mile section of railway called the *Joint Line* is the artery through which most Powder River Basin coal reaches the rest of the United States. It is the busiest stretch of railroad in the world, handling 60 mile-long coal trains a day.[6] Moving the same volume of coal by truck—currently the only alternative to rail—is both prohibitively expensive and restricted by available trucks and drivers. Rail is the only practical way to transport this critical resource.

The destruction of an important bridge, like the High Triple Bridge over Antelope Creek, would stop coal transport on one of two primary lines feeding rail hubs for distribution to multiple states. Destruction of one to three similar targets immediately before peak periods of seasonal electricity demand could disable much of the country's generation capacity for periods of weeks to months.

Powder River Basin coal illustrates one of the most vital interdependencies among CIKR sectors. A normal accident to the railway network could magnify and spread to fuel shortages in the Midwest. These shortages could reduce or halt the generation of electric power to much of the United States. Blackouts or brownouts might then propagate to more populous states like the 2003 blackout did, causing millions of people to go without power. If the coal shortage

[4]http://www.eia.gov/

[5]https://en.wikipedia.org/wiki/BNSF_Railway

[6]http://www.sourcewatch.org/index.php?title=Powder_River_Basin#Rail_history

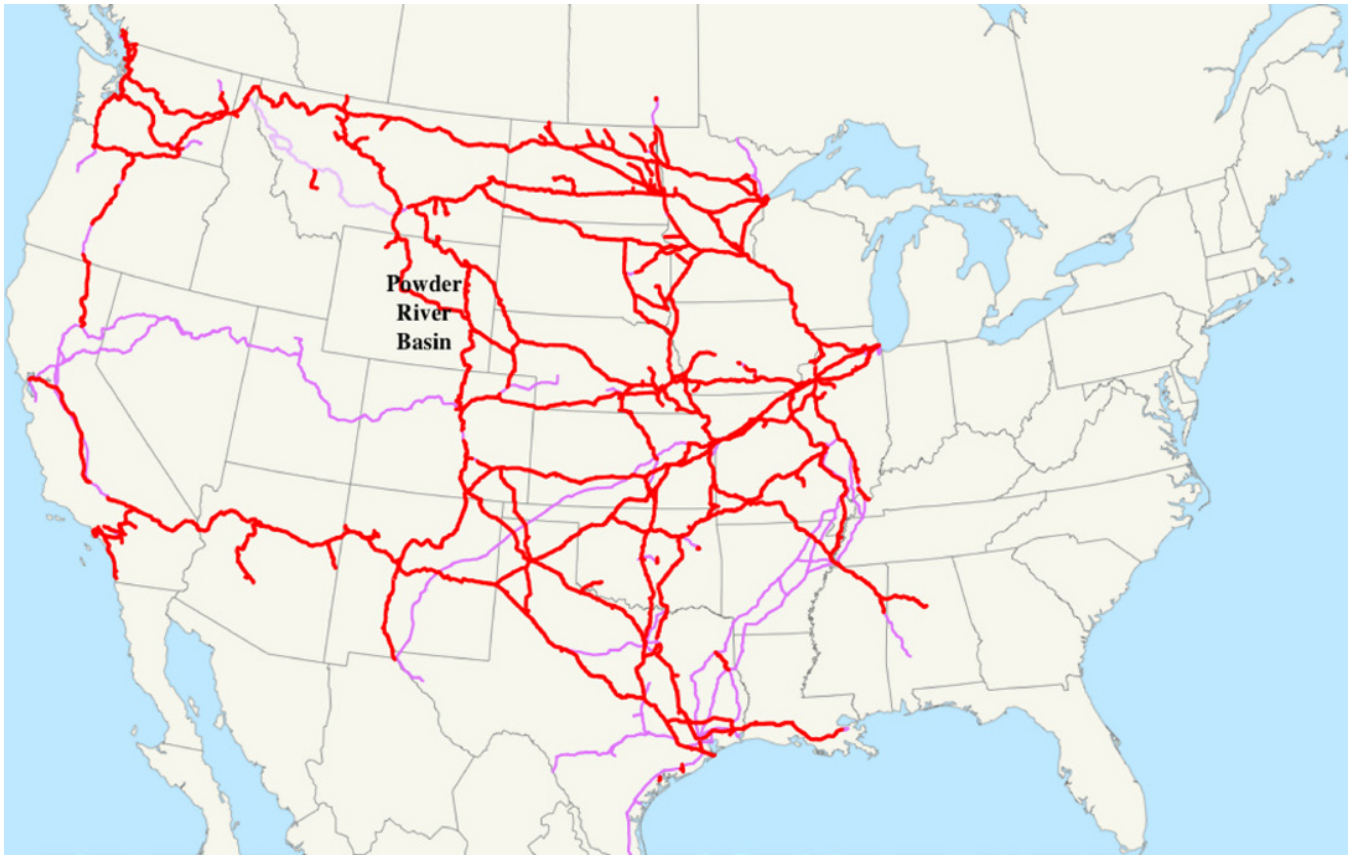


FIGURE 12.3 BNSF railway network connects the largest source of coal in the United States to power plants throughout the Midwest and markets abroad (Asia) via the Columbia River Gorge between Oregon and Washington. This is a file from the Wikimedia Commons: https://en.wikipedia.org/wiki/File:BNSF_Railway_system_map.svg.

lasted long enough, the cascade failure could more broadly damage the economy of the entire country.

12.4 THE RISE OF OIL AND THE AUTOMOBILE

When Edwin Drake discovered petroleum in Titusville, Pennsylvania, in 1859, the market for oil was to make kerosene for illuminating homes and offices.[7] Pennsylvania soon dominated the kerosene lamp oil business, holding 80% market share. But then two important things happened: (1) Edison perfected the electric light bulb, and (2) Henry Ford revolutionized transportation with the mass-produced gasoline-powered automobile. The Edison electric light bulb was better and cheaper than a kerosene lamp, and the automobile was a better form of transportation than the horse and buggy. Both transformed the fledgling oil industry from a niche business to a mainstream consumer products economy. What was not so evident in the 1860s was the extent to which oil would become dependent on transportation, and vice versa.

[7]http://www.pipeline101.com/History/index.html

12.4.1 Oil

The oil business quickly became intertwined with transportation because the well that produced kerosene was in Pennsylvania and the kerosene and gasoline consumer lived miles away in New York. Getting the product from wellhead to market became the tail that wagged the dog. In the early days, crude oil was poured into wooden whiskey barrels and carried by wagon from the wellhead to the train station where it was transported by rail to Northeast refineries. After refining it was once again distributed by horse-drawn wagons to consumers in New York City and other Northeastern cities. Moving oil became as profitable as the oil itself—a problem that John D. Rockefeller quickly solved by making an exclusive deal with the railroad company.

The teamsters controlled the supply lines from wellhead to consumer. And like any monopoly, they began charging high prices for moving crude to refinery and refined kerosene to consumer. The cost to move a whiskey barrel of oil 5 miles to a rail station was greater than the railroads charged to move the same barrel from Pennsylvania to New York City. An alternative to this chokehold had to be found. Thus was born the first pipeline in 1863. Pipelines

were soon, and still are today, the most economical way to move petroleum from wellhead to refinery and from refinery to consumer.

The famous Tidewater pipeline, opened in 1879 to bypass even more costly middlemen, was the first *trunk* line—major energy transmission pipeline. It was half-owned by John D. Rockefeller who secured control of the oil supply chain in order to control the oil market. Thus monopoly power passed from the teamsters to Rockefeller. He rapidly expanded the pipeline network to Buffalo, Cleveland, and New York.

The business model for petroleum and NG was well established when Texas oil was discovered. Extract raw crude from the ground, send it to the Northeastern refineries by transmission pipeline, and then distribute the refined product to large metropolitan centers in the East. Thus a profitable energy business was linked to control of a very long transportation network that often exceeded 10,000 miles in length. The barrier to entry was significant—how many competitors could afford to build such a pipeline? It soon became a monopoly—Standard Oil—run by Rockefeller.

Standard Oil was not the only monopoly held by a single powerful industrialist in the late 1800s. But it was one of the most persistent. The Sherman Anti-Trust Act was passed in 1890, making certain practices of a monopoly illegal. The Hepburn Act of 1905 declared transborder transmission lines a form of interstate commerce and hence subject to regulation by the federal government. But it was not until 1912 that Standard Oil was forced to break up after a lengthy and colorful struggle involving President Theodore Roosevelt and powerful industrialists of the Gilded Age. In the end, seven regional companies replaced Rockefeller's Standard Oil Company, as the petroleum industry continued to grow at a rapid pace. After World War I, pipeline networks were serving much of the nation, rising to 115,000 miles in the 1920s. Today they exceed 200,000 miles.

12.4.2 Natural Gas

NG and its more compact form, LNG, are also transported largely by pipeline. Pipeline middlemen controlled the market and hence the price of getting NG to consumers. So in 1938 Congress passed the NGA, which allowed the FPC (forerunner of FERC) to set the prices charged by interstate pipelines, but not the prices charged by producers.[8] In 1940 the NGA was amended to add regulation of the NG facilities themselves to FPC's responsibilities.

In the famous *Phillips decision* of 1954, the US Supreme Court determined that the NGA covered wellhead prices as well. The FPC was now responsible for regulating the prices along much of the supply chain. The Court sought to protect consumers from "exploitation at the hands of natural gas companies."[9] The government sought to control the source, refining, and distribution of gas and oil for the benefit of taxpayers and voters.

Oil production in the United States was outstripped by demand during the 1950s and 1960s, which led to an increasing dependence on imported oil. The *Colonial Pipeline* constructed in 1968 to deliver oil products from the Gulf of Mexico states to the Northeastern United States was the largest privately financed project in history up to that time. The 800 mile Trans-Alaska pipeline delivered 2 million barrels/day when it opened in 1977. It delivers 1 million barrels today.

Increasing demand and decreasing domestic supply pushed importation of oil ever upward. By 2003, the United States was consuming 20 million barrels of oil *per day*, 68% of it delivered by pipeline networks and 27% by boat. The remainder was moved by truck and rail. Most NG and oil is moved from wellhead to refinery to distributor, and then to consumer, by pipe. This is the most economical way to provide an enormous quantity of product over such long distances to so many consumers. But it is also the source of vulnerabilities, as we shall see later in Section 12.6.2.

The following network analysis of one major pipeline system connecting Canadian sources with US refineries and consumers indicates a low spectral radius, but a high betweenness centrality similar to pipeline networks in other CIKR. Like most flow networks, energy supply chain networks are characterized by low spectral radius and high betweenness centrality, which makes them vulnerable to denial-of-service failures.

12.5 ENERGY SUPPLY CHAINS

Supply chain management (SCM) is the management of all steps in the delivery of a product or service to consumers. The gas and oil industry is principally a SCM commons consisting of exploration, drilling, operation of crude pipelines, and operation of refineries for the production of fuels, plas-

[8]There seems to be a pattern here: federal government regulators attempted to put limits on the prices that an electric utility may charge consumers but allowed the wholesale prices of power to float. This backfired in states such as California because of shortages in the capacity of the power grid to distribute power to consumers. In the case of NG, regulation had to be expanded to cover the entire supply chain. Is this the eventual fate of electric power grid operators?

[9]Phillips Petroleum Co. v. Wisconsin, 1954. This Supreme Court decision resulted in an expansion of the FPC's jurisdiction, and in the aftermath of the decision, natural gas applications under the Natural Gas Act exploded, far exceeding the volume of electrical and hydroelectric regulation handled by the FPC.

tics, and so on. Trunk line or "transmission pipes" are the arterials that deliver refined products such as gasoline and aviation fuel to terminals located around the country. *Distribution* refers to the sale and delivery of these products to consumers from storage terminals.

A greatly simplified supply chain model is shown in Figure 12.4. Unrefined product from a wellhead is transported to a refinery by a pipeline system or boat. The refinery converts the crude into refined products, which in turn are transmitted over long distance to terminals, where they are stored—typically in large tanks. Then a distribution network of trains, pipelines, and trucks delivers the product to consumers. Transmission generally refers to the long-haul pipes and distribution to the local short-haul delivery to consumers.

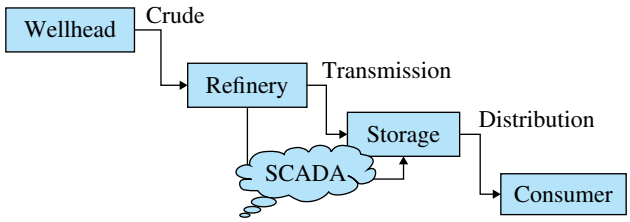


FIGURE 12.4 Simplified supply chain model of NG and petroleum energy infrastructure.

The vast miles of pipeline are monitored by various SCADA systems that report anomalies such as leaks and broken components. An example of a pipeline SCADA network is given in the chapter on SCADA—the crude pipeline that delivers oil from the fields around Bakersfield, California, to the refineries in the Los Angeles area. SCADA is an increasingly important part of the supply chain because of environmental and security concerns.

12.5.1 PADDs

For purposes of tracking supply and demand reporting, the supply chain is divided into five regions called *Petroleum Administration for Defense Districts* (PADDs) (see Fig. 12.5). They were created during World War II when gasoline was rationed. The division is somewhat artificial today, but still used to record inflows and outflows of petroleum and petroleum products. Broadly speaking, the five PADDs cover the East Coast (PADD1), the Midwest (PADD2), the Gulf Coast (PADD3), the Rocky Mountain Region (PADD4), and the West Coast (PADD5). Supply chain data can be obtained from the US Department of Energy's Energy Information Administration, which collects and publishes oil supply data by PADD.[10]

[10]www.eia.doe.gov

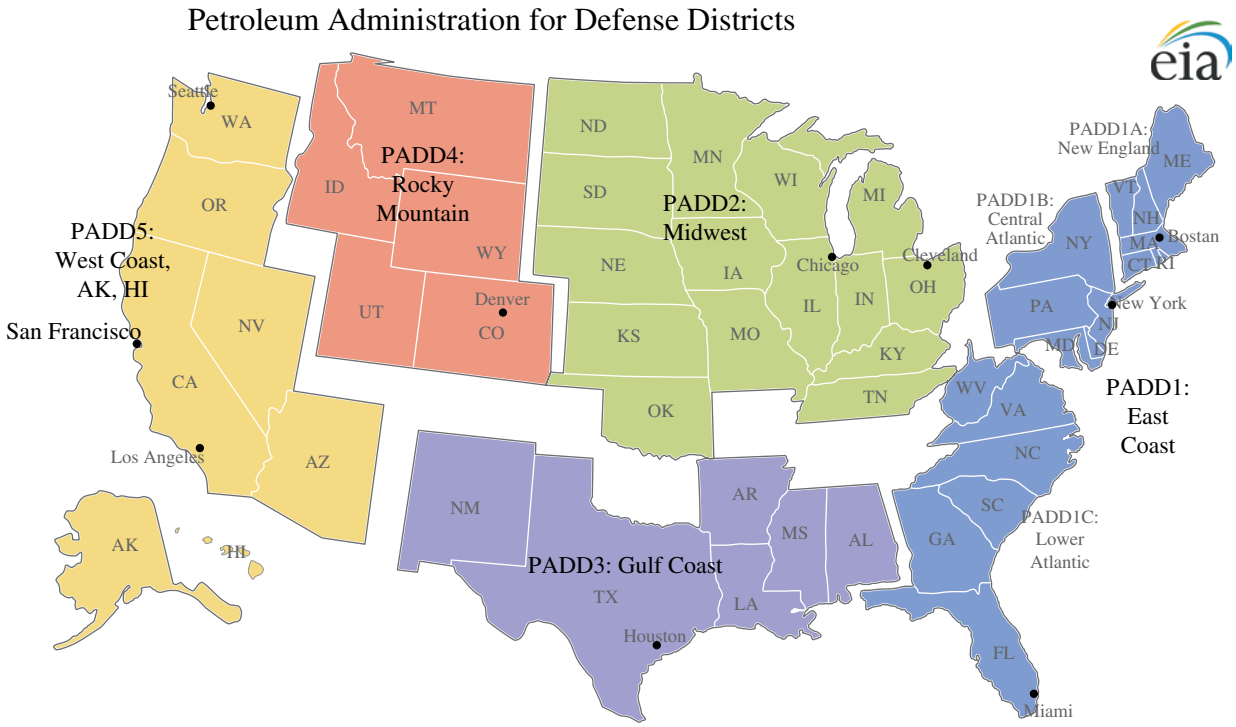


FIGURE 12.5 Petroleum Administration for Defense Districts (PADDs) are still used today to track oil production and consumption. http://www.eia.gov/.

For example, in 2001, the percentage of oil produced and imported by each PADD was:

PADD #	Production (%)	Importation (%)
1. East Coast	~0	~100
2. Midwest	10	90
3. Gulf Coast	90	10
4. Rocky Mountain	~100	~0
5. West Coast	45	55

Today oil flows mainly from Canada or the Gulf Coast to refineries in PADD3, and refined product flows mostly from PADD3 to PADD1 and PADD2 (East Coast and Midwest). The West Coast supply chain is almost a stand-alone network that depends on Canadian, Alaskan, and foreign sources.

12.5.2 Refineries

The United States has more refining capacity (20%) than any nation in the world and refines 96% of all petroleum products it uses. Refineries are distillation factories. That is, they take in crude oil and distill it into various petroleum products according to their specific gravity (density). Crude oil is separated into a variety of petroleum products because lighter molecules percolate to the top of the distillery and heavier particles stay near the bottom.

A 42 gallon barrel of crude oil produces 44 gallons of refined product because of gains in the distillation process. Less than one-half is turned into gasoline. Assuming your automobile tank holds 20 gallons, each of the 140 million registered automobiles and trucks in the United States consumes 140 million barrels of oil just to "fill 'er up"! (In 2001 US consumers used 840 million gallons/day. Filling up all registered vehicles consumed at least one-sixth of total consumption for 1 day.)

The products obtained from a barrel of oil are summarized below along with the amount of each. Note that one barrel of oil is roughly equal to one tank of gasoline for a typical passenger automobile:

Gallons	Product
19.5	Gasoline
9.2	Heating oil
4.1	Jet fuel
11.2	Asphalt, feedstock, lubricants, kerosene, and so on
44	Total

In 2003 there were 152 refineries in 32 states. The top 10 refineries produced almost 20% of the total, and the top 2 (Baytown, TX, and Baton Rouge, LA), produced over 5% of the national supply. Refineries are highly concentrated along the Gulf Coast. In fact, most of the high-volume refineries are clustered along the coastline between Galveston, TX and Baton Rouge, LA. This high concentration poses a major vulnerability in the refinery component of the energy supply chain.

12.5.3 Transmission

Large pumps and compressors move billions of gallons of product along pipelines each year. It is not uncommon for a pumping station to be powered by a 4000 hp diesel or electric motor, for example. While these pumps and compressors are backed up with auxiliary power, it shows how dependent the energy sector is on the power sector. Ironically, without power, energy cannot be moved along its supply chain. And without energy, power cannot be generated.

In addition, operators depend on SCADA to monitor the pipeline and its contents—looking for leaks and other anomalies. SCADA is less critical, but without it, operators are blind. Writing for the Allegro Energy Group, Cheryl Trench describes how pipeline SCADA works:

> Pipeline employees using computers remotely control the pumps and other aspects of pipeline operations. Pipeline control rooms utilize Supervisory Control And Data Acquisition (SCADA) systems that return real-time information about the rate of flow, the pressure, the speed and other characteristics. Both computers and trained operators evaluate the information continuously. Most pipelines are operated and monitored 365 days a year, 24 hours per day. In addition, instruments return real-time information about certain specifications of the product being shipped—the specific gravity, the flash point and the density, for example—information that are important to product quality maintenance. Oil moves through pipelines at speeds of 3 to 8 miles per hour. Pipeline transport speed is dependent upon the diameter of the pipe, the pressure under which the oil is being transported, and other factors such as the topography of the terrain and the viscosity of the oil being transported. At 3–8 mph it takes 14 to 22 days to move oil from Houston, Texas to New York City. [2]

Figure 12.6 illustrates how petroleum products are "sequenced" in a pipeline. Like packets of data on the Internet, multiple segments travel on the same pipe. For example, a segment of kerosene may be transported along with a segment of gasoline. This leads to *transmixing*—the unintentional mixing of products, which often requires some

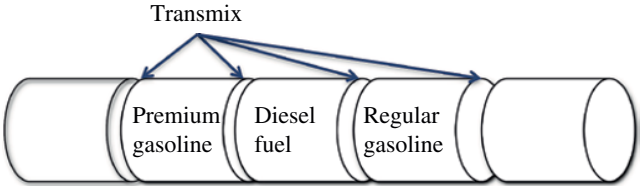


FIGURE 12.6 Pipelines are "multiplexed" by combining different products on the same pipeline according to their density.

reprocessing at a terminal. But this is an extremely efficient way to move different products over the same (expensive) network. Allegro Energy Group president, Cheryl Trench, describes how sequencing works (Colonial Pipeline is the largest oil pipeline in the United States):

> Pipeline operators establish the batch schedules well in advance. A shipper desiring to move product from the Gulf Coast to New York Harbor knows months ahead the dates on which Colonial will be injecting heating oil, for instance, into the line from a given location. On a trunk line, a shipper must normally "nominate" volumes—ask for space on the line—on a monthly schedule… As common carriers, oil pipelines cannot refuse space to any shipper that meets their published conditions of service. If shippers nominate more volumes than the line can carry, the pipeline operator allocates space in a non-discriminatory manner, usually on a *pro rata* basis. This is often referred to in the industry as "apportionment." (Pages 15–16 in Ref. [2])

12.5.4 Transport4

In August 1999 several major pipeline companies formed a joint venture to create Transport4 (T4)—a SCM Web site dedicated to scheduling product sequences through the major pipelines. Buckeye Pipe Line Company, Colonial Pipeline, Explorer Pipeline, and TEPPCO Pipeline opened up T4 to connect any pipeline carrier to any customer. Today 80% of all petroleum products are scheduled via T4.

Advanced technologies such as horizontal drilling and hydraulic fracturing (*fracking*) have opened up new oil fields, which in turn have outstripped pipeline reach and capacity. Fracking uses water to break apart rocks to release oil deposits. Horizontal drilling has made the Bakken Oil Play—an area spanning Alberta, Canada, North Dakota, and parts of Wyoming—the second largest oil field in North America, with North Dakota second only to Texas in terms of crude oil production.

The opening of new oilfields has had a dramatic impact on freight rail transportation, in regions where pipelines do not exist to transport the crude to refineries. The slack has been taken up by railroads that transport barrels of oil from the North and West to refineries in the East and South. Newfound sources of oil in Canada and the United States have outpaced building of pipelines. The *KeystoneXL* proposal to enhance pipeline reach and capacity is studied later in Section 12.9.

12.5.5 Storage

Products transmitted through major pipeline systems like the 5500 mile Transcontinental Pipeline (Transco) that delivers 95 million gallons/day to the East Coast (PADD1) end up in storage farms where millions of gallons of heating oil, aviation fuel, and gasoline are stored prior to being distributed to consumers. These tanks are large, conspicuous, and concentrated in a few critical places. As a consequence, they contribute enormously to supply chain risk. This topic is addressed in Section 12.6.3.

12.5.6 Natural Gas Supply Chains

The NG network architecture is much like the petroleum network architecture. It is characteristically subject to low spectral radius, high betweenness centrality, and geographical concentration of critical assets. In some ways, it is even more critical because NG currently provides 23% of US energy and the percentage is rising. Thus its criticality is growing.

The NG supply chain is even more expansive than the petroleum supply chain. It consists of 280,000 miles of transmission pipeline and 1.4 million miles of distribution pipeline. Furthermore, it is expanding at a rate of 14%/decade. Over the next 20 years, the United States will need 255,000 more miles of pipeline.[11] The National Petroleum Council estimates that utilities will spend $5 billion/year to build out this capacity.

12.5.7 SCADA

SCADA systems are designed to keep pipeline systems safe. They do this by monitoring pumps, valves, pressure, density, and temperature of the contents of the pipeline. An alarm sounds when one or more of the measurements go out of bounds, so operators can shut down pumps and compressors. Operators must consider the local terrain, the product that is inside a pipe, and numerous physical characteristics of the pipeline. If safety limits are exceeded, a SCADA system can automatically shut down a pipeline within minutes.

As we shall see, SCADA plays a relatively minor role in vulnerability analysis because a SCADA shutdown may cause loss of revenue, but not loss of life. Also, SCADA vulnerabilities can be mitigated for a relatively modest investment compared with the investment required to replace a refinery, storage tank, or section of pipeline.

12.6 THE CRITICAL GULF OF MEXICO CLUSTER

Refineries, major transmission pipelines, and major storage facilities are the most critical assets in the energy sector because of their large capacities and geographical concentration. In addition, many of these critical components are wide open—they are easily accessed and therefore at risk due to symmetric and asymmetric attacks, weather-related damage, and industrial accidents.

[11]National Petroleum Council: Meeting the Challenge of the Nation's Growing Natural Gas Demand, December 1999.

Gas and oil supply chains are rich targets for many hazards from nature and humans. The most obvious hazards are weather related, especially along the hurricane-battered Gulf Coast states. This analysis is limited to the most frequent threat–asset pairs:

Major refineries along the Gulf Coast
- Refinery–equipment failure (ref–equipment)
- Refinery–human error (ref–human)
- Refinery–miscellaneous accidents (ref–misc)

Major transmission pipelines serving the Northeast market
- Transmission–equipment failure (trans–equipment)
- Transmission–corrosion (trans–corrosion)
- Transmission–aging (trans–age)

Major storage facilities serving the Northeast market
- Storage–lightning/static electricity (store–lightning)
- Storage–operational accidents (store–operation)
- Storage–leaks and ruptures (store–leak)

Figure 12.7 shows the two major networks containing critical assets. The Gulf of Mexico oil fields where exploration and drilling occur and the refineries that turn crude oil into refined petroleum products feed petroleum products into the massive Colonial pipeline network running north from Houma, Louisiana, to Linden Station, New Jersey. The (also massive) storage complex at Linden Station stockpiles gasoline and commercial airliner fuel until it is needed. Disruption of any link in this supply chain can have severe consequences because it supplies most of the energy consumed by the major metropolitan areas in the Northeast.

12.6.1 Refineries

The Gulf Coast area south of Houston, TX, stretching east to Lake Charles, LA, and then to Baton Rouge, LA, contains 5 of the top 10 refineries in the nation. These 10 produce nearly 20% of all refined petroleum products for the nation:

1. Baytown, TX
2. Baton Rouge, LA
3. Texas City, TX
4. Whiting, IN
5. Beaumont, TX
6. Deer Park, TX
7. Philadelphia, PA
8. Pascagoula, MS
9. Lake Charles, LA
10. Wood River, IL

The five largest refineries located along the critical Gulf Coast region produce 11% of the total US supply. Table 12.2 lists the top five and how many barrels they produce each day. The largest refinery in the United States belongs to ExxonMobil—located in Baytown, TX (in the Galveston Bay south of Houston), and surrounded by smaller refineries as well. It produces aviation fuels, lubricants, base stocks, chemicals, and marine fuels and lubricants.

In 1900, Galveston was the site of the deadliest hurricane in US history. Over 8000 people lost their life. Then in 1947, the largest port disaster in US history wiped out Texas City. A ship containing fertilizer caught on fire and spread throughout the port and much of the town. Over the years, the worst refinery disasters have occurred in large clusters like Texas City: Whiting, Indiana; Texas City, Pasadena, and Amarillo, Texas; Baton Rouge, Louisiana; Romeoville, Illinois; and Avon and Torrance, California. The bottom line is that concentrations such as the Galveston Bay cluster are both vulnerable to damage and highly critical to the oil supply chain.

Refineries can be shut down because of fires, lack of power, or weather-related damage. Power outages may last for only a few hours. Destruction of crude oil pipelines can deny service to a refinery for perhaps days. An explosion or fire can cause longer-term damages—perhaps for months. Hurricanes can flood pipelines and refineries, rendering them unproductive. The cost and likelihood of each incident varies and may be difficult to estimate when performing a risk analysis.

Figure 12.8 lists the most frequent causes of refinery accidents and shutdowns. The top three are related to equipment failures (25%), human error (24%), and miscellaneous (22%) incidents such as fire, falls, hazardous material spills, and electrical malfunctions.[12] For example, 17 refineries in Louisiana reported 301 accidents in 2011. Refinery accidents can cost hundreds of millions of dollars.

The cost of closing a refinery depends on the volume of output (lost revenues) and the size of the refinery. Refinery replacement can cost more than $1 billion, and the loss of production (500,000 barrels/day) can have severe implications on revenues as well as shortages that lead to price increases at the gasoline station. The economic impact of faults in this supply chain is inestimable.

12.6.2 Transmission Pipelines

According to PHMSA, there is over 2.3 million miles of transmission and distribution pipelines in the United States carrying NG, petroleum, and refined petroleum products, as well as chemicals and hydrogen. Crude oil pipes transport unrefined oil from wellhead or ship to refinery. Product pipelines move refined products from refinery to storage facilities at the head end of distribution networks. Distribution

[12]www.SouthernStudies.org data for 17 refineries in Louisiana during 2005–2011.

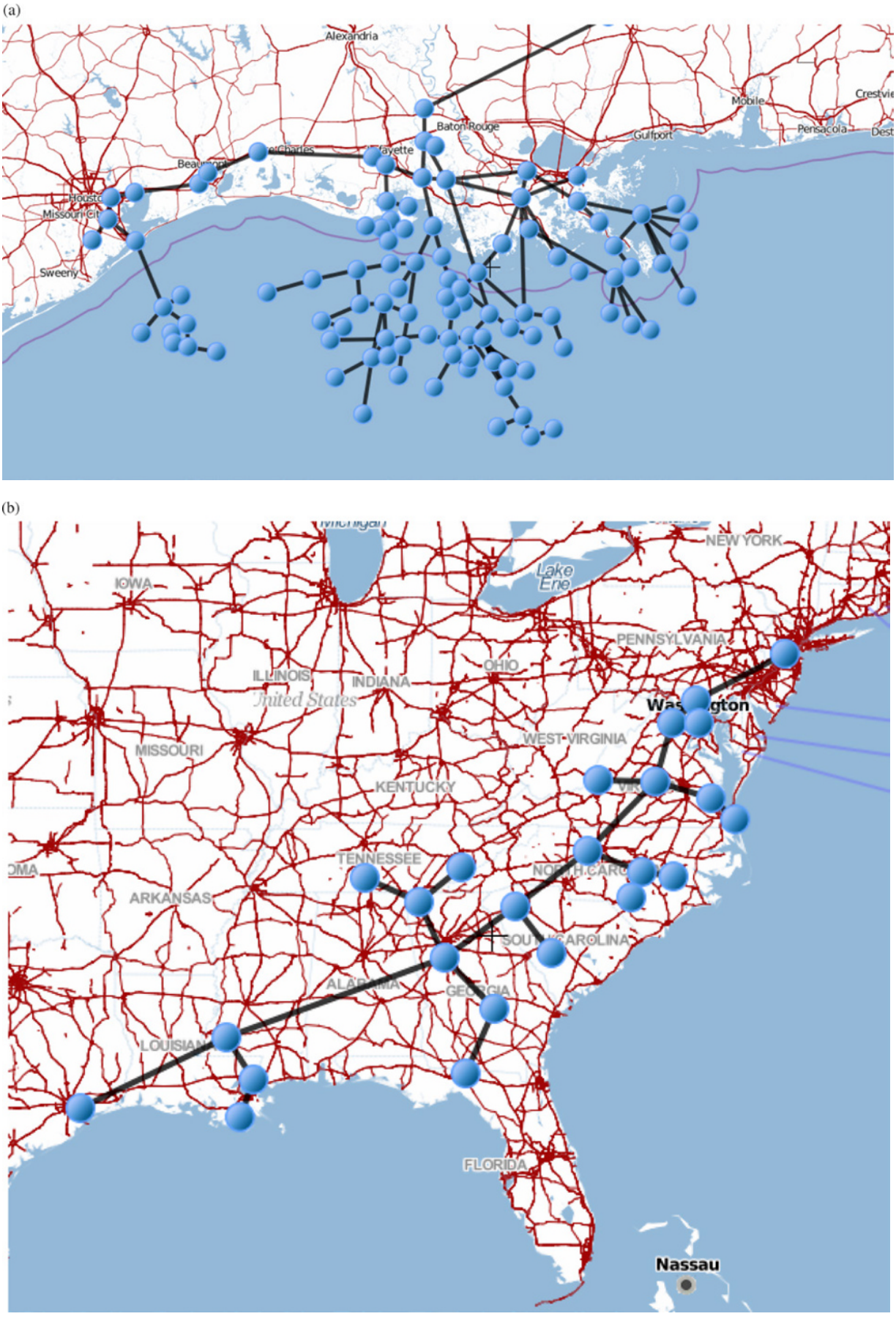


FIGURE 12.7 Refined petroleum products flow from the Gulf of Mexico oil field and LOOP terminal to refineries along the Gulf Coast and then to consumers in the Northeast by way of the Colonial pipeline and storage network. (a) Gulf of Mexico network of oil fields and refineries. LOOP is circled. (b) Colonial pipeline connecting refineries to markets in the Northeast. http://www.shellpipeline.com/cd_maps/SPL403_D_gc_crude_f.pdf.

TABLE 12.2 Rank, name, and location of the most productive refineries in the Gulf of Mexico region produce 11% of national refined product (2.2 of 20 MMbl/day)

Rank#	1	2	3	5	9
Corporation	Exxon Mobil	Exxon Mobil	BP PLC	Exxon Mobil	PDV AMERICA
Refiner	ExxonMobil	ExxonMobil	BP Products	ExxonMobil	Citgo Petroleum
Location	Baytown, TX	Baton Rouge, LA	Texas City, TX	Beaumont, TX	Lake Charles, LA
# of barrels/day	523,000	491,500	437,000	348,500	32,4300
Market (%)	11.0				

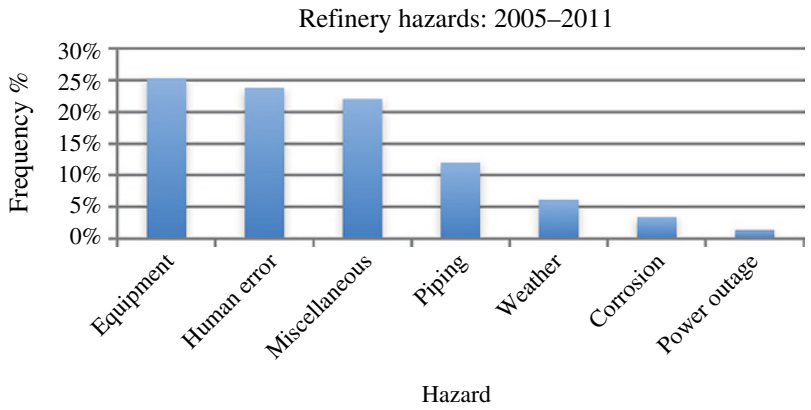


FIGURE 12.8 Equipment failure and human error are the top refinery hazards.

networks complete the delivery of refined products to consumers.

The largest gas and oil transmission pipelines are owned and operated by conglomerates such as Kinder-Morgan, Colonial Pipeline, Transco, and Keystone. For example, the 5500 mile-long Colonial Pipeline transmits 95 MMbl/day of petroleum products to customers in PADD1, eventually ending up in Perth Amboy, New Jersey, where it is stored before being distributed by Buckeye Pipeline. It delivers gasoline, kerosene, home heating oil, diesel fuels, and national defense fuels to shipper terminals in 12 states and the District of Columbia. It transports 20% of the entire national supply of refined petroleum products.

Colonial is owned by a joint venture among several major energy companies:

Koch Capital Investments Co.	25.27%
HUTTS LLC	23.44%
Shell Pipeline Co. LP	16.12%
CITGO Pipeline Investment Co.	15.8%
Phillips Petroleum International Investment Co.	8.02%
Conoco Pipe Line Co.	8.53%
Marathon Oil Co.	2.82%

Figure 12.9 shows the top hazards affecting both gas and oil pipelines. Equipment failures, corrosion, operational accidents, aging, and miscellaneous accidents are the top causes of pipeline spillage. The largest spills can cost upwards of $85 million, but most are much smaller. The fractal dimension of consequences for gas pipeline accidents is approximately 1.0, which means risk is at its tipping point between low and high risk. The fractal dimension of consequences for oil pipelines is approximately 0.50, which means oil pipelines are high-risk networks. Fortunately, fatalities and injuries from pipeline accidents have been declining since 1970.

12.6.3 Storage

Storage tanks hold LNG and refined petroleum products while they are waiting to be distributed through a network of jobbers and resellers. For example, Cushing, OK, is the site of the largest crude oil storage facility in the world, with a capacity of 61 million barrels. The US-DOE *Strategic Petroleum Reserve* (SPR) can hold up to 727 million barrels in underground caverns spread around the Southeastern United States. The nation's largest NG storage site is an underground cavern near Cedar Creek Field, Montana, with a capacity of 287 billion cubic feet. Linden Station, located at Perth Amboy, NJ, is the largest refined product storage facility in the Northeastern United States and the terminal point for the Gulf of Mexico supply chain.

In 1996, Colonial Pipeline Co. delivered 820.1 million gallons of jet fuel directly to airports in PADD1: Dulles International Airport, Baltimore–Washington International

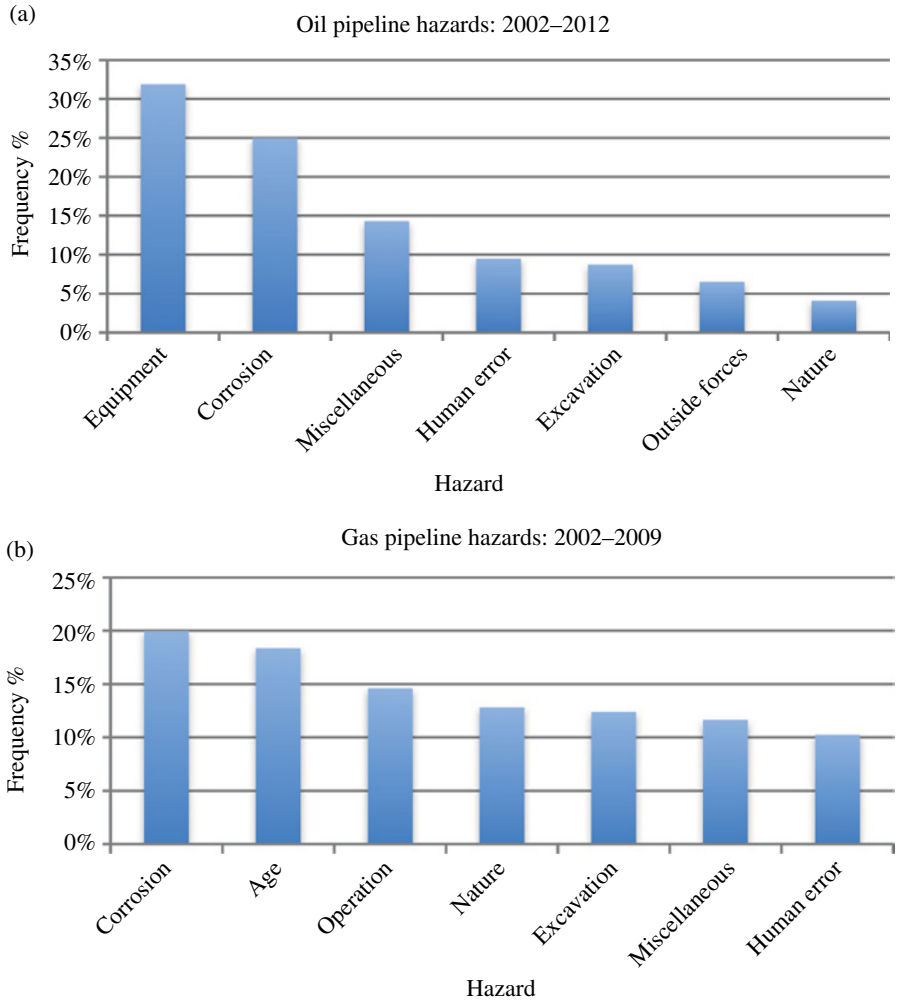


FIGURE 12.9 Equipment failure, corrosion, operational accidents, and aging are the top gas and oil pipeline hazards. (a) The top oil pipeline hazards are equipment failure, corrosion, and miscellaneous accidents. (b) The top gas pipeline hazards are corrosion and operational accidents.

Airport, Nashville Metropolitan Airport, Charlotte Douglas Airport, Raleigh–Durham Airport, Greensboro Triad, and Hartsfield–Atlanta International Airport. Thus, the transportation sector depends on this CIKR.

Colonial and other common carriers terminate at Perth Amboy, New Jersey, in the New York harbor. This storage terminal contains a cluster of storage tanks that are obvious supply chain vulnerabilities. While this analysis is focused on the Linden Station, note the numerous other storage facilities in Figure 12.11. Hundreds of storage tanks are located in close proximity to one another providing a high-value target to terrorists and criminals.

On January 2, 1990, a ruptured pipeline at Linden Station spilled an estimated 567,000 gallons of fuel oil into the Arthur Kill waterway between New Jersey and Staten Island. The spill caused extensive environmental damage and economic consequences. Pumping continued for 9 h after the rupture because the owner/operator (Exxon) was slow to detect the spill—operators had disabled the leak detection system. A Coast Guard team in a small boat saw oil bubbling to the surface and determined that the pipeline was the source after operators conducted a pressure test by pumping more oil into the water.

In 1996, storage tank faults were the second most significant cause of spillage after pipelines.[13] Generally, the most frequent hazards affecting storage tanks are lightning strikes and static electricity (38%), operational/reaction accidents (27%), and leaks/ruptures (13%) [3] (see Fig. 12.10). A ranked exceedence probability distribution study of the largest storage faults between 1963 and 2002 produced a fractal dimension of 1.14. This short-tailed exceedence probability suggests that storage failures are low risk. However, the con-

[13]www.pipelinesafetyfoundation.org

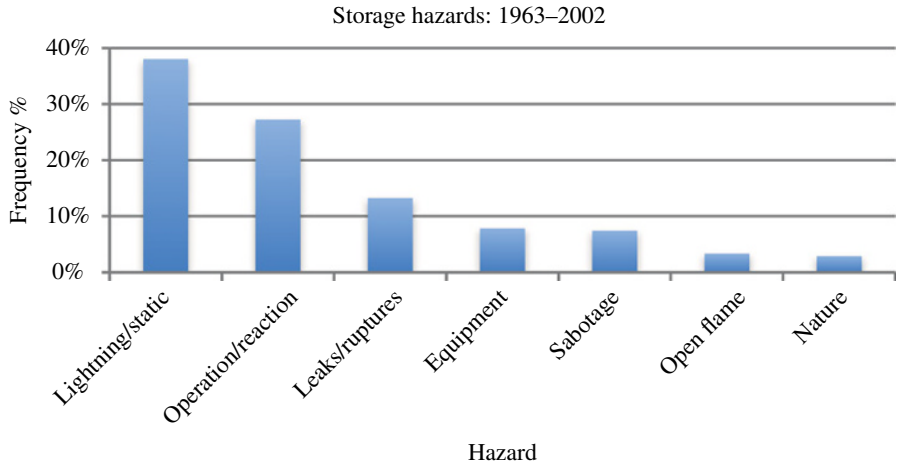


FIGURE 12.10 Lightning/static electricity, operational/reaction accidents, and leaks/ruptures are the top storage hazards.

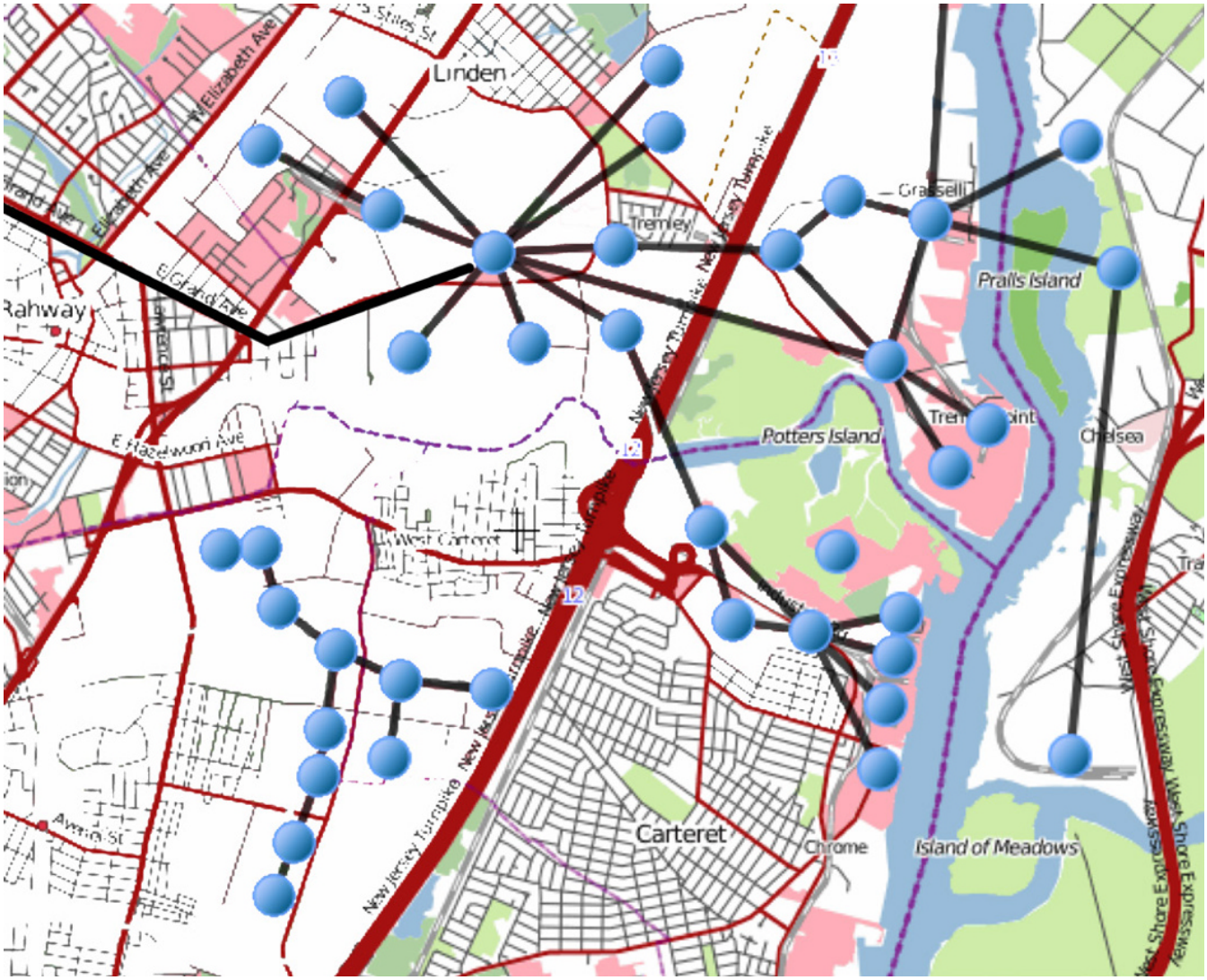


FIGURE 12.11 Linden Station is the focal point of the massive storage facility at the end of the Colonial Pipeline supply chain. The major storage facilities and pipelines are shown here as a network straddling the New Jersey Turnpike.

sequences of the largest failures can exceed $250 million. Fortunately, they are very rare.

12.7 THREAT ANALYSIS OF THE GULF OF MEXICO SUPPLY CHAIN

Figure 12.12a contains a general fault tree for the energy sector and its major components at risk—refineries, transmission, and storage. This fault tree is applied to the Gulf of Mexico supply chain shown in Figure 12.7. Hypothetical values are used to protect the supply chain's security, but it is obvious from the foregoing analysis that it will fail if one or more refinery, transmission pipeline, or storage facility fails. There is little to no redundancy in this CIKR network.

The three major asset types are represented in Figure 12.12a by three components. The refinery component faces three threats—equipment failure (ref–equipment), human error (ref–human), and miscellaneous accidents (ref–misc). The transmission component faces three threats—equipment failure (trans–equipment), corrosion (trans–corrosion), and aging (trans–age). Finally, the storage component's three most likely hazards are lightning/static electricity (store–lightning), operational accidents (store–operation), and leaks/ruptures (store–leak). These threat–asset pairs are represented in Figure 12.12a along with probabilities obtained from data presented in Figures 12.9, 12.10, and 12.11.

For simplicity, consequences from all hazards are assumed to be $2000 million, and all threats are set to 100%. Thus, initial vulnerability and elimination costs are the only variables

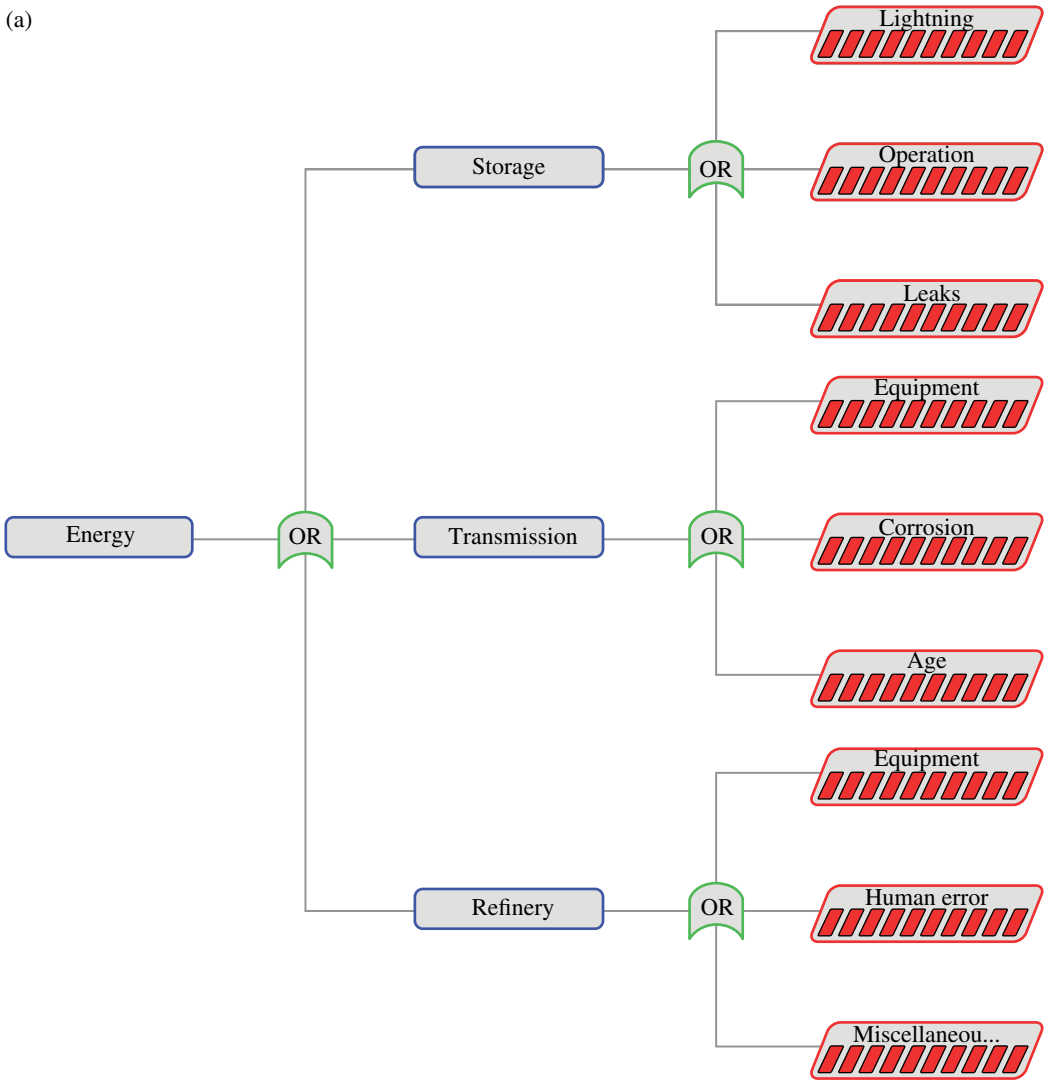


FIGURE 12.12 General fault tree model of probable threat–asset pairs in the energy supply chain for three major components: refineries, pipelines, and storage facilities. (a) General fault tree for the energy sector contains threat–asset pairs for critical assets in the supply chain.

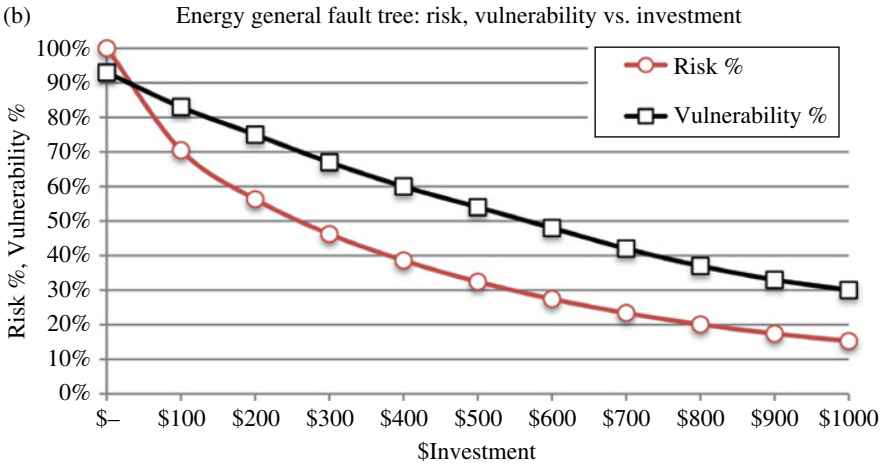


FIGURE 12.12 (Continued) (b). Risk and vulnerability versus investment for hypothetical values plugged into the general fault tree show that vulnerability is more difficult to reduce than risk.

TABLE 12.3 Allocation of $300 million reduces energy fault tree risk from $4484 million (100%) to $2073 million (46%) for hypothetical input data

Name	Threat (%)	Vulnerability (%)	Elimination cost $(millions)	Consequence $(millions)	Risk initial	Allocation $(millions)	Vulnerability reduced (%)	Risk reduced
Lightning	100.00	38.00	100.00	2000.00	760.00	53.23	5.48	109.63
Operation	100.00	27.00	150.00	2000.00	540.00	50.46	8.91	178.18
Leaks	100.00	13.00	1000.00	2000.00	260.00	0.00	13.00	260.00
Equipment	100.00	31.90	200.00	2000.00	638.00	61.04	11.09	221.74
Corrosion	100.00	25.00	250.00	2000.00	500.00	40.78	14.79	295.76
Equipment	100.00	25.00	300.00	2000.00	500.00	28.19	18.47	369.49
Human error	100.00	24.00	50.00	2000.00	480.00	32.20	3.10	62.02
Miscellaneous	100.00	22.00	400.00	2000.00	440.00	0.00	22.00	440.00
Age	100.00	18.30	100.00	2000.00	366.00	34.10	6.79	135.84

in Table 12.3. Initial risk is $4484 million and initial vulnerability (probability of a supply chain failure) is 93%.

Figure 12.12b suggests that fault tree vulnerability is more difficult to reduce than risk. Once again, this is due to the OR-gate logic in the model. If any one or combination of threat–asset pairs fails, the entire supply chain fails. Risk falls below 50% after approximately $300 million is invested, so this number will be used in the following analysis.

If the risk ranking investment strategy is used to allocate $300 million, the threat–asset pairs would be funded in order: lightning–store (R = $760 million), equipment–trans (R = $638 million), and operation–store (R = $540 million). However, this is nonoptimal because of the differences in initial vulnerability and elimination costs. An optimal allocation strategy reduces the vulnerability of threat–asset pairs in different order: equipment–trans (allocation = $61 million), lightning–store (allocation = $53 million), and operation–store (allocation = $50 million). This illustrates an important difference in strategies. Is the objective to reduce the worst-case risk or overall risk? Risk minimization reduces overall risk.

Optimal risk reduction can be a harsh strategy. For example, two threat–asset pairs are denied any investment at all: leak–store and human–refinery pairs receive zero funding. Why? These two have the highest elimination cost to consequence ratios: leak–store = 0.50 and human–ref = 0.20. All other threat–asset pairs range from 0.13 down to 0.03. Risk minimization ranks threat–asset pairs according to the return on investment. Lower ratios mean risk reduction is less expensive and therefore preferred over more expensive risk reductions.

12.8 NETWORK ANALYSIS OF THE GULF OF MEXICO SUPPLY CHAIN

PADD3 (Gulf of Mexico) oil fields form a major network of refineries, pipelines, and a major import port called Louisiana Offshore Oil Port (LOOP). In the foregoing introduction, refineries were shown to be critical because the nation's largest are nodes in this network. In addition, LOOP accounts for approximately 13% of the nation's total import of crude.

Taken all together, the PADD3 network in Figure 12.7 is vital to the energy sector.

Network analysis of the Gulf oil field network shows that it is self-organized around both hubs and high betweenness nodes and links. Its fundamental resilience line indicates high resilience against cascade failures, but low robustness of pipeline links. Deeper analysis of the Gulf oil field network shown in Figure 12.7 identifies its critical nodes and links, assuming betweenness and connectivity centrality are the most important properties. This network has 106 nodes, 121 links, and a mean connectivity of 2.28 links per node. Its spectral radius is 4.11, or 1.8 times mean connectivity. This modest amount of self-organization means cascade failures are less important than flow failures. When connectivity and betweenness centrality are combined, self-organization becomes apparent because of relatively high betweenness centrality.

Link robustness is very low at (121–106)/121 = 12%, but node robustness is much better at (1–1/4.11) = 75.6%. Node robustness is high because many terminal (wellhead) nodes lie under the ocean in the Gulf of Mexico. There are 42 blocking nodes (40%) that hold this portion of the Gulf of Mexico oil supply chain together. They lie on a path established by nodes and links connecting the Gulf oil fields and refineries to Linden Station.

This critical path contains critical links and nodes (SS, ST, MC, and JE nodes are under the water):

Critical nodes:

Houma, Gibson, SS-28, ST-300, and Clovelly

Critical links:

SS-28 → Gibson, MC-311-Nairn, Erath → Gibson, West Columbia → East Houston, and JE → Gibson

A subset of these critical nodes and links form a *critical path* leading to Capline and the Colonial Pipeline (see Fig. 12.13). A fault in any one of these nodes or links leads to a denial of oil to the Colonial Pipeline and therefore to the markets in the Northeastern United States. Summarizing, the critical nodes and links along this critical path are:

Critical nodes along critical path:

Houma, Gibson, Clovelly, St. James, Calpine

Critical links long critical path:

Houma → Gibson, SS-28 → Gibson, SS-332 → Houma, JE → Houma, Erath → Gibson

Simulation of flow through this core network produces a high-risk flow exceedence probability distribution with fractal dimension approximately 0.37. Therefore, this network is extremely fragile. Failure in one critical node or link can reduce output as much as 60%, assuming the hypothetical values used by the simulation. This, of course, is due to the very high betweenness centrality and low link robustness, which leads to a very critical path from wellhead to refinery to storage tank.

To harden the entire Gulf oil field network, 40% of all nodes would have to be 100% hardened. This may be prohibitively expensive and practically impossible. Without a redundant network of pipelines and refineries, the Gulf of Mexico supply chain will remain vulnerable to natural or human-made disruptions.

12.9 THE KEYSTONEXL PIPELINE CONTROVERSY

The Cushing Oil Trading Hub (COTH) in Cushing, Oklahoma, is a dramatic example of *Gause's law*. Oil entrepreneurs rushed to the Glenpool Basin area 15 miles south of Tulsa following discovery in 1905 of a major oil gusher. Oklahoma quickly became the nation's largest oil producer as other fields were discovered and exploited. By 1914, the Cushing field was producing 50,000 barrels/day, or one-quarter of the entire state's production. Cushing became a center for exploration and production of nearby oil fields. In 1928, the Oklahoma City Field was discovered and soon became the nation's largest oil producing basin. Major oil companies followed, including Sinclair Oil, Marland Oil (merged with Conoco in 1929), Cities Services Oil Company (CITGO), Phillips Petroleum Co., American Association of Petroleum Geologists, Halliburton, Noble Corporation, Anderson & Kerr Drilling (Kerr-McGee, purchased by Anadarko Petroleum in 2005), and others.

As oil fields began to run dry in the 1940s, production became less important and brokering became more important. COTH became the official price settlement point for the West Texas Intermediate (WTI) crude—the pricing benchmark for North American crude. It evolved into a storage depot and an oil-trading center handling crude oil from Gulf of Mexico imports and domestic sources to Midwest refining markets. As supplies from domestic sources declined, the crude oil pipeline system was reversed to transmit crude oil produced in Alberta, Canada, to the COTH where it could be distributed to both Midwest and Gulf Coast refineries. More recently, other crude pipelines reversed flow direction due to increasing domestic supplies from shale reserves in North Dakota and Texas, bringing even more crude oil through COTH. COTH has become the nation's hub for crude oil.

The COTH receives imported Canadian crude via the Keystone pipeline (590,000 barrels/day) for distribution to refineries. A proposed KeystoneXL pipeline from Canada to the COTH and then to Port Arthur, Texas, will increase Keystone Pipeline System capacity to 1.3 million barrels/day. The Obama administration denied construction of the pipeline on environmental concerns, but a study done by

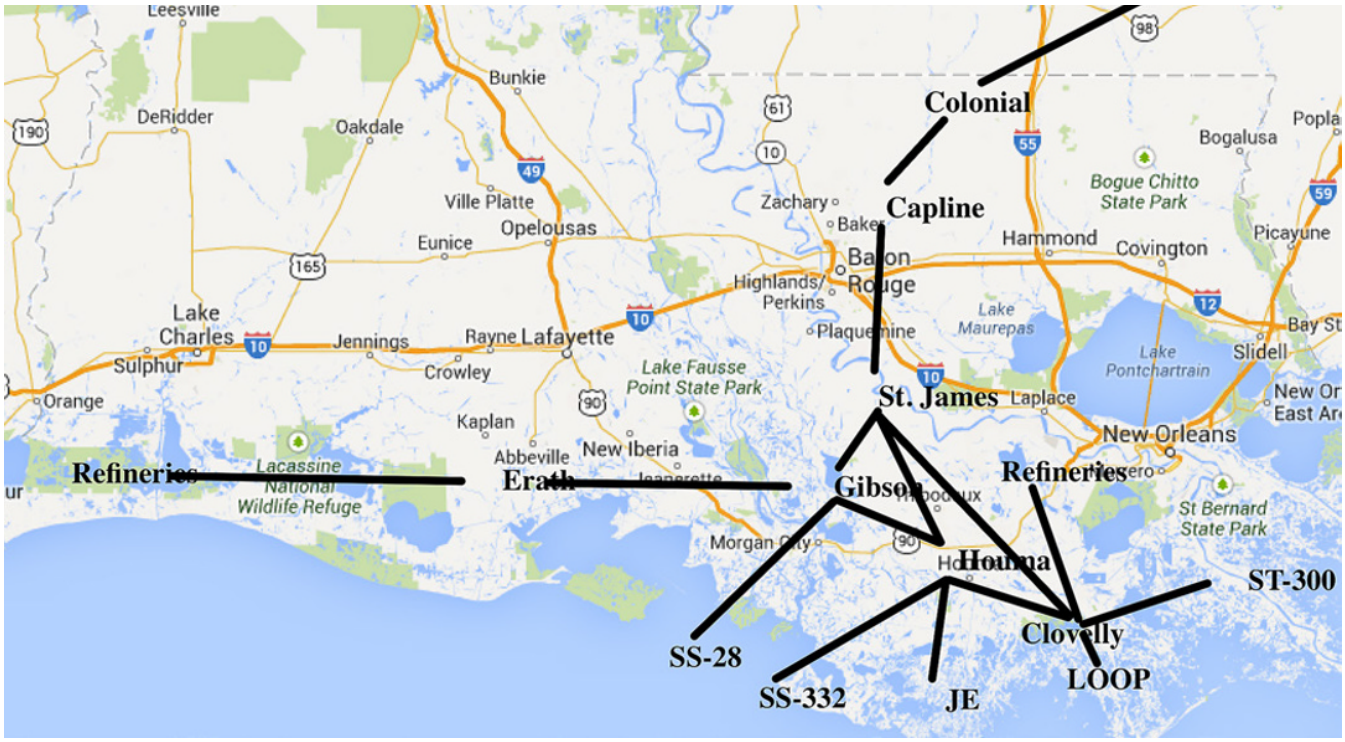


FIGURE 12.13 The core of the Gulf of Mexico oil field network is centered on Houma and Gibson in Louisiana.

Larrañaga *et al*. concluded that the addition of KeystoneXL increases resilience of this supply chain by 55% [4]. The KeystoneXL analysis is reported in Chapter 4.

The KeystoneXL case illustrates the dramatic impact that policy and regulation has on CIKR. In fact, regulation is the major factor shaping most interstate infrastructure. It is responsible, along with economic concerns, for creating hubs, betweeners, and dangerous concentrations of critical assets. Self-organized criticality of the energy sector is mainly due to economic and political factors.

12.10 THE NATURAL GAS SUPPLY CHAIN

The NG (methane) supply chain is very similar to the oil supply chain. NG is gathered from crude oil wells, separated from water, oil, and other contaminants, and pumped into NG pipelines. It is compressed and pumped along vast pipeline networks at roughly 30 miles/h, typically traveling 10,000 miles or more in a matter of days before reaching storage facilities near consumer markets. It is compactly stored in liquid form (LNG) by cooling it to minus 260 connectivities and then thawed out before distributing to consumers. Large-capacity LNG ships also take advantage of the compactness of gas while in transit between producers and consumers.

SCADA plays an important role in the safe transmission of NG by regulating the pressure that moves the gas along miles of pipeline. Compressor stations are located about every 50–100 miles to regulate speed and pressure. SCADA provides the necessary control information to regulate supplies as consumer demand varies. By compressing the gas, output can be sped up or slowed down.

The NG supply chain also mimics the oil supply chain in terms of vulnerabilities and risk. Clustering of pipelines in the same geographical location reduces operational and supply costs, but it also increases the risk of attack or accidental damage because of clustering, lack of redundancy, and high betweenness centrality in transmission networks. In addition, environmental regulations and NIMBY (Not In My Back Yard) play predominant roles in self-organization of NG networks and storage clusters.

Most NG comes from the Gulf of Mexico coast and Canada and heads toward the East Coast where the large metropolitan populations consume large quantities of NG to heat homes (80%) and generate electrical power (20%). The largest of these transmission networks is the Transcontinental Pipeline, also known as Transco. Transco runs from Houston to the New York harbor terminal (see Fig. 12.14).

Table 12.4 lists the largest NG pipeline networks along with their capacities and pipeline lengths. All of the pipelines are over 10,000 miles long, and taken together, they

FIGURE 12.14 The 10,600 mile Transco Pipeline carries liquid natural gas (LNG) from fields along the Gulf of Mexico and ports along the East Coast to markets in the Northeast.

TABLE 12.4 The largest NG pipelines are more than 10,000 miles long and move over one-third of all NG in the nation

Rank	1	2	3	4	5	6	7
Name	Transcontinental	Columbia	Tennessee	ANR	Texas Eastern	Dominion	El Paso
Owner	Williams	NiSource	El Paso	El Paso	Duke	Dominion	El Paso
Capacity (MMbl/day)	7,362	7,276	7,271	6,667	6,438	6,275	4,882
Length (miles)	10,636	11,215	14,761	10,600	12,118	10,000	10,200

account for 35% of all NG in the United States. The Transco pipeline pumps nearly 10 billion cubic feet (Bcf)/day, and the longest, Tennessee pipeline (14,700 miles), pumps nearly as much. Together, these 7 networks account for 79,500 of the 212,000 miles (38%) of pipeline and collectively pump 46.8 MMbl/day of the nation's 133 MMbl/day consumption.

Three networks listed in Table 12.4 are of particular interest because they provide 16% of the national supply and nearly all of the NG energy consumed by the populous PADD1 region (Eastern United States). Number 1 Transco, number 3 Columbia, and number 6 Dominion form a critical node near Cove Point, Maryland. The *Cove Point Intersection* illustrates a typical risk in energy supply chains.

West of Washington DC, across the Potomac River in the Virginia counties of Fairfax, Prince William, and Loudoun lies the nexus of three major pipelines: Transco, Columbia, and Dominion (aka CNG). Dominion runs through Leesburg Station; Transco runs through Loudoun Station, Nokesville, and Dranesville. Pleasant Valley forms a network node through which 16% of the nation's supply of NG flows. This node connects LNG pipelines from Dominion, Transco, and Columbia lines. Disruption in this geographic region could interrupt 20,000 MMbl/day of LNG and NG on its way to New York and points north. Cove Point Intersection is a critical node in the NG supply chain because of its potentially high consequence.

12.11 ANALYSIS

The energy supply chain is vast and complex. Coal resources are concentrated in just a few geographical locations in the United States and distributed through massive railway networks throughout the country. While coal is a cheap and plentiful fuel, its use is being restricted by environmental rules that limit CO_2 emissions. Regulation is the biggest threat to this asset and will limit its future use in power generation.

The other major sources of energy—gas and oil—depend on critical supply chains that run all the way from domestic wells in the Gulf of Mexico and North Dakota, as well as foreign wells in Canada and Saudi Arabia, to the homes and offices of every American. These vast oil and NG networks are characterized by geographical clustering of refineries, single-point-of-failure pipelines, and extremely large-capacity and concentrated storage terminals. These targets are extremely attractive because they are also extremely consequential.

Oil and NG supply chains are regulated much like the electric power grid and are also undergoing radical transformation due to deregulation of their industries and environmental regulation. Existing pipelines are long and capable of hauling enormous quantities of energy. But they are subject to the competitive exclusion principle, which means the entire nation depends on only a few *hub carriers* like Colonial Pipeline, Transco, Dominion Pipeline, and Colombia Pipeline to carry most of the supply. The uniqueness of these *common carriers* translates into high-consequence targets for both accidents and terrorists. We have no choice but to protect these assets, because building redundant capacity is not economically feasible. Thus, oil and NG supply chains will continue to be highly structured, low link redundant, high-risk networks.

Energy supply chains are not likely to be enhanced in any significant way for decades. According to the 2001 national energy policy,

> There are over two million miles of oil pipelines in the United States and they are the principal mode for transporting oil and petroleum products. Virtually all natural gas in the United States is moved via pipeline. Pipelines are less flexible than other forms of transport, because they are fixed assets that cannot easily be adjusted to changes in supply and demand. Once built, they are an efficient way to move products. A modest sized pipeline carries the equivalent of 720 tanker truckloads a day—the equivalent of a truckload leaving every two minutes, 24 hours a day, 7 days a week.[14]

The topology of this CIKR sector has evolved for over a century. It is unlikely to restructure in less than another century. Even so, environmental regulation and growing interest in renewable sources of energy will gradually reshape this sector over the next 100 years. With so much at stake, the only feasible strategy is to protect what we have and respond to accidents and attacks on critical nodes and links while transitioning to next-century supply chains.

[14] http://www.whitehouse.gov/news/releases/2001/06/energyinit.html

12.12 EXERCISES

1. Which of the following is a measure of energy?
   a. Horsepower
   b. kWh
   c. kW
   d. Watts
   e. Gallons
2. Most energy consumed in the United States comes from:
   a. Natural gas
   b. Petroleum
   c. Coal
   d. Hydroelectric
   e. Wind and solar
3. Gas and oil pipeline safety is monitored by:
   a. FERC
   b. NERC
   c. EPA
   d. OPS
   e. DHS
4. The energy sector is regulated by:
   a. FERC
   b. NERC
   c. EPA
   d. OPS
   e. DHS
5. The largest transmission pipeline, and largest privately financed project in the United States at the time, is:
   a. Colonial Pipeline
   b. Transco Pipeline
   c. Kinder-Morgan Pipeline
   d. East Texas Gas Pipeline
   e. Dominion Pipeline
6. Most pipelines are monitored and controlled by:
   a. OPS
   b. SCADA
   c. E-ISAC
   d. FERC
   e. DHS
7. How many PADDs are there in the United States?
   a. 5
   b. 2
   c. 3
   d. 1
   e. 50 (one per state)

8. In the United States, the top 10 refineries produce:
   a. Nearly all of the gas and oil
   b. 20% of all petroleum products
   c. 5% of all petroleum products
   d. Over half of all gas and oil products
   e. Two-thirds of all gas and oil products
9. The energy sector began deregulation in:
   a. Energy Policy Act of 1992
   b. PURPA in 1978
   c. HLPSA of 1979
   d. FERC Order 436 in 1985
   e. DHS in 2003
10. The Gulf of Mexico oil supply chain is characterized by:
    a. Congestion
    b. High redundancy of transmission
    c. Low link robustness
    d. Low node robustness
    e. High spectral radius
11. Powder River Basin coal is vulnerable to:
    a. Earthquakes
    b. EPA
    c. Competition
    d. Power outages
    e. Railway disruptions
12. The largest NG transmission pipeline for the PADD1 area (Northeastern United States) is:
    a. Transco
    b. Columbia
    c. Dominion
    d. Tennessee Pipeline
    e. Colonial Pipeline
13. The largest refined petroleum product pipeline in the United States is:
    a. Transco
    b. Columbia
    c. Dominion
    d. Tennessee Pipeline
    e. Colonial Pipeline
14. The top refinery hazards in recent years have been:
    a. Terrorism
    b. Flooding
    c. Corrosion
    d. SCADA failure
    e. Equipment failure
15. The top oil pipeline hazards in recent years have been:
    a. Terrorism
    b. Flooding
    c. Corrosion
    d. SCADA
    e. Equipment failure
16. The top gas pipeline hazards in recent years have been:
    a. Corrosion
    b. Terrorism
    c. Equipment failure
    d. Human error
    e. Hurricanes
17. The top storage hazards in recent years have been:
    a. SCADA faults
    b. Equipment failure
    c. Hurricanes
    d. Lightning/static electricity
    e. Regulation
18. Why is the Cove Point Intersection at risk?
    a. Watson is the source or entry point to the pipeline.
    b. Fault probabilities are too high.
    c. It is the intersection of three large pipelines.
    d. It is near the capital.
    e. Casualties would be high.
19. Why is Linden Station at risk?
    a. Consequences of corrosion would be high.
    b. It is near New York City.
    c. It is near New Jersey.
    d. It has large storage tanks.
    e. It is near water.
20. The most likely factor shaping the energy sector in the future is:
    a. The United States will run out of goal.
    b. Environmental regulation will limit coal.
    c. Solar and wind will replace natural gas and oil.
    d. The United States will run out of gas and oil.
    e. Nuclear power will become number one.

12.13 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. How will introduction of mass-market electrical vehicles impact the energy budget of the United States as shown in Figure 12.1?

B. Why are pipelines generally resilient against cascade failures but fragile in terms of flow resilience and link robustness? Suggest a way to increase link robustness in the Colonial pipeline network.

C. Why is the Gulf of Mexico cluster considered highly critical? Explain your reason in terms of redundancy and network science.

D. Why is gas and oil pipeline safety the responsibility of the Department of Transportation instead of the Department of Energy?

E. Over half of all energy is lost between production and consumption. Suggest ways that this sector might reduce this high loss rate.

REFERENCES

[1] U.S. Energy Information Administration, *Annual Energy Review 2008*, DOE/EIA-0384(2008), Washington, DC, June 2009. Projections: AEO2010 National Energy Modeling System, run AEO2010R.D111809A.

[2] Trench, C. J. *How Pipelines Make the Oil Market Work—Their Networks, Operation and Regulation*. A memorandum prepared for the Association of Oil Pipe Lines and the American Petroleum Institute's Pipeline Committee, December 2001. Available at www.aopl.org. Accessed June 30, 2014.

[3] Chang, J. I. and Lin, C.-C. A Study of Storage Tank Accidents, *Journal of Loss Prevention in the Process Industries*, 19, 2006, pp. 51–59.

[4] Smith, P. K., Bennett, J. M., Darken, R. P., Lewis, T. G., and Larrañaga, M. D. Network-Based Risk Assessment of the U.S. Crude Pipeline Infrastructure, *International Journal of Critical Infrastructures*, 10, 2013, pp. 67.

13

ELECTRIC POWER

In 2000, the National Academy of Engineering named modern power grids—those vast electrical power generation, transmission, and distribution networks that span the country—the top engineering technology of the twentieth century. In the Academy's opinion, the power grid surpassed invention of the automobile, airplane, moon shot, atomic bomb, delivery of safe and abundant water, and electronics as the most important engineering accomplishment. Electrical power is what makes modern society tick. It is essential. So it comes as no surprise that the grid is one of the fundamental infrastructures of the United States.

In this chapter you will learn the following concepts and be able to apply them to the challenge of electrical power grid risk analysis:

1. *Blackouts are increasing*: The frequency and size of power outages increased rapidly after deregulation in 1992 but has leveled off since 2012. The high risk of power outages is traced to a number of factors, including, but not limited to, underinvestment in transmission and distribution, deregulation of utilities resulting in loss of control, and network topology—rising self-organized criticality (SOC) due to the grid's wiring diagram.
2. *Deregulated utilities*: Historically, the components of power—generation, transmission, distribution, load (consumption), and SCADA control—have been owned and operated by *vertically integrated* utility companies. Since 1992 these vertically integrated monopolies have been disaggregated and decoupled from generation, transmission, and distribution of power through deregulation legislation. Unfortunately, deregulation has brought with it *economic and control vulnerabilities* that are still being worked out. By separating key components of the grid into competing companies, regulation has introduced instabilities in command and control of the grid.
3. *Deregulation and physics*: The power grid has been and continues to be shaped by a combination of governmental regulation and the laws of physics—these two do not always work together. Physics demands rigorous control of complex electrical circuits. Deregulation often ignores this requirement by separating control from the operators, thus introducing instability. A deregulated grid is like a highway network with thousands of vehicles going in different directions: accidents are bound to happen.
4. *There is no shortage of power*, but there is a shortage of transmission and distribution capacity. The United States produces approximately 15% more power than it consumes. But it cannot always deliver power to where it is needed, when it is needed, because of inadequate transmission and distribution capacity. This occasionally leads to blackouts—massive normal accidents that start small and spread to far points of the grid.
5. *Criticality of power plants*: No single power generator is critical—the largest source of power provides less than 1% of the national capacity. It is a myth that the most vital components of the nation's power sector are

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

power plants. This points once again to the "middle" of the grid as the most likely place for failures to occur.

6. *Topology matters*: grid topology magnifies the frequency and size of blackouts. The grid has comparatively low spectral radius, so it should be resilient against cascade failures, but it is self-organized around high-betweenness critical paths. These critical paths exist along transmission lines connecting power plants and major population centers. Resiliency can only be improved by reducing this form of self-organization.
7. *Resiliency and congestion*: The Western power grid (WECC96) is used to illustrate the relationship between critical points called hot spots and congestion. It is shown that high betweenness and connectivity are correlated with known congestion points in the WECC96. Congestion cannot be removed by increasing the capacity of one or more transmission line. Instead, the grid must be rewired to lower betweenness. Resiliency and congestion are related to network topology.
8. *Human threats*: Major threat–asset pairs are traced to fuel supply chains, destabilizing physical and cyber attacks, and attacks on critical components such as transformers and major transmission lines. Threat analysis shows that it is easier to reduce risk than vulnerability, because of the OR-gate relationship in the fault tree model of terrorist threats.
9. *Distributed generation*: An alternative topology that solves many of the problems facing the grid is distributed generation—colocating power generation near its load. This can be achieved in two ways: (1) by switching to solar, wind, or alternative sources of power or (2) by adding storage to the grid. Both of these reduce reliance on long-haul transmission lines.

13.1 THE GRID

The grid, as the collection of electric power networks across the country is called, is a complex CIKR system consisting of four major components as shown in Figure 13.1. Power is *generated* by burning a fossil fuel or turning a turbine by wind, water, or tidal action and then put into a vast *transmission*

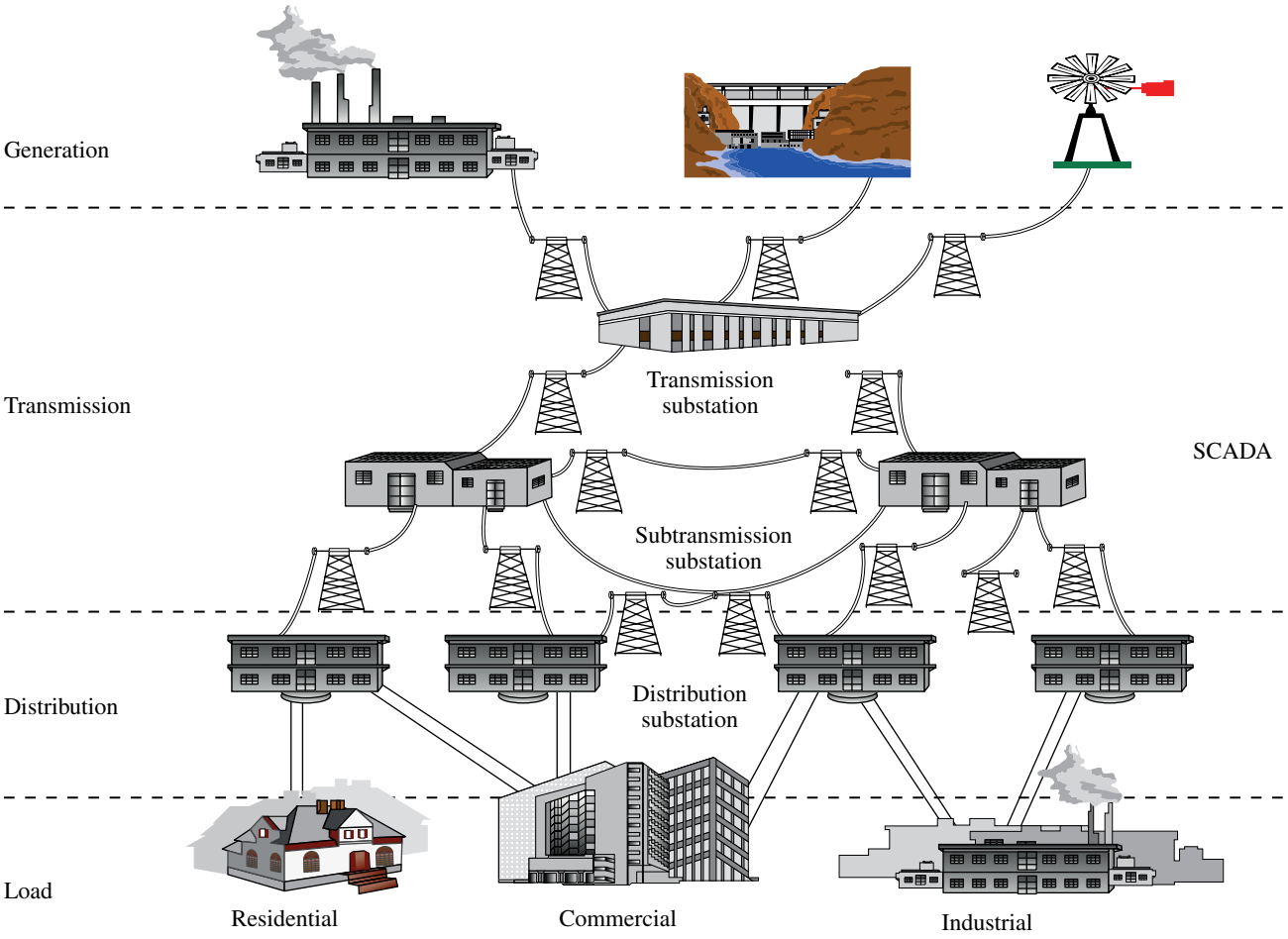


FIGURE 13.1 The five major components of the power grid are generation, transmission, distribution, load, and ICS-SCADA, typically called an energy management system (EMS).

system that transports the electrons long distances to local *distribution* networks. Transmission is more economical if done at high voltages, so it must be stepped down and redirected by substations along the way. Metropolitan-level distribution networks further distribute stepped-down power to residential, commercial, and industrial customers called the *load*.

The entire *generation*, *transmission*, and *distribution* process is monitored by an industrial control *SCADA*, typically called an energy management system. This ICS-SCADA system is typically out of band, which means it is a separate communications network running parallel to the Internet and power lines, although this practice is changing as SmartGrid technologies are adopted. *SmartGrid* is the name given by Massoud Amin to describe the convergence of electric power SCADA with information technology and communications [1]:[1]

> Amin's idea came from years of work on stability of aircraft and other complex systems. He was impressed by a pilot who landed a jet fighter after one wing was blown off and control surfaces were damaged. The clever pilot used thrust control to land the badly damaged airplane. This got Amin to thinking: if a pilot can control a badly damaged aircraft, then it ought to be possible for a computer to control the unruly electric power grid. To do so, operators need real-time information about the state of the grid at any instant in time. The challenge is to complete the loop from power plant to transmission, distribution, load, and back—quickly enough to head off impending disaster… computers connected to embedded sensors in power lines, transformers and electricity meters can be programmed to overcome instabilities created by unpredictable faults in the network.

The "unruly electric power grid" became increasingly unruly following deregulation in 1996 due to a number of economic and regulatory missteps. Department of Energy data suggests that at a national level, the number of outages and their size peaked around 2012 and has been declining (see Fig. 13.2). Figure 13.2b shows a relatively high PML risk during the period 2000–2014. However, risk remains, and the power grid faces new stability challenges with the transition from burning fossil fuel to renewable fuels such as solar and wind. The national grid still suffers from a shortage of transmission lines and storage. Figure 3.6 documents the decline in investment since reaching a peak in the early 1970s.

At a national level, power grid risk and resilience appear to be improving, but regional outages still occur with regularity. The extremely long tail of Figure 13.2b suggests room for additional improvement. In addition, the rise of electric vehicles, increased electrification due to the build-out of large data processing centers, and exponential increase in the Internet of Things (IoT) indicates continued pressure on this sector.

13.2 FROM DEATH RAYS TO VERTICAL INTEGRATION

The Grid has its historical roots in the famous Pearl Street New York utility created by Thomas Edison in the 1880s. This first utility supplied direct current (DC) electrical power to 59 Manhattan customers. Edison was convinced that DC was the best way to deliver electricity, but Serbian immigrant Nikola Tesla had a better idea: alternating current (AC). Tesla was Edison's rival in all things having to do with harnessing the power of the electron. He is the father of all modern electric power generation technology (generators), distribution (transmission lines and substations), and appliances (motors).

A titanic power struggle between Tesla and Edison ensued over the advantages of AC versus DC. When Tesla sold his patent rights to George Westinghouse, Edison's feud shifted from Tesla to Westinghouse. Edison derided AC. At one point he used the electric chair to convince consumers that AC was unsafe. Tesla countered with daring demonstrations of his own:

> Tesla gave exhibitions in his laboratory in which he lighted lamps without wires by allowing electricity to flow through his body, to allay fears of alternating current. He was often invited to lecture at home and abroad. The Tesla coil, which he invented in 1891, is widely used today in radio and television sets and other electronic equipment. That year also marked the date of Tesla's United States citizenship.
>
> Westinghouse used Tesla's system to light the World's Columbian Exposition at Chicago in 1893. His success was a factor in winning him the contract to install the first power machinery at Niagara Falls, which bore Tesla's name and patent numbers. The project carried power to Buffalo by 1896.
>
> In 1898 Tesla announced his invention of a tele-automatic boat guided by remote control. When skepticism was voiced, Tesla proved his claims for it before a crowd in Madison Square Garden.
>
> In Colorado Springs, Colo., where he stayed from May 1899 until early 1900, Tesla made what he regarded as his most important discovery—terrestrial stationary waves. By this discovery he proved that the Earth could be used as a conductor and would be as responsive as a tuning fork to electrical vibrations of a certain frequency. He also lighted 200 lamps without wires from a distance of 25 miles (40 kilometers) and created man-made lightning, producing flashes measuring 135 feet (41 meters). At one time he was certain he had received signals from another planet in his Colorado laboratory, a claim that was met with derision in some scientific journals.
>
> Tesla was a godsend to reporters who sought sensational copy but a problem to editors who were uncertain how seriously his futuristic prophecies should be regarded. Caustic criticism greeted his speculations concerning communication with other planets, his assertions that he could split the Earth like an apple, and his claim of having invented a death ray capable of destroying 10,000 airplanes at a distance of 250 miles (400 kilometers).[2]

[1] A personal communication with Massoud Amin.

[2] http://www.neuronet.pitt.edu/~bogdan/tesla/bio.htm

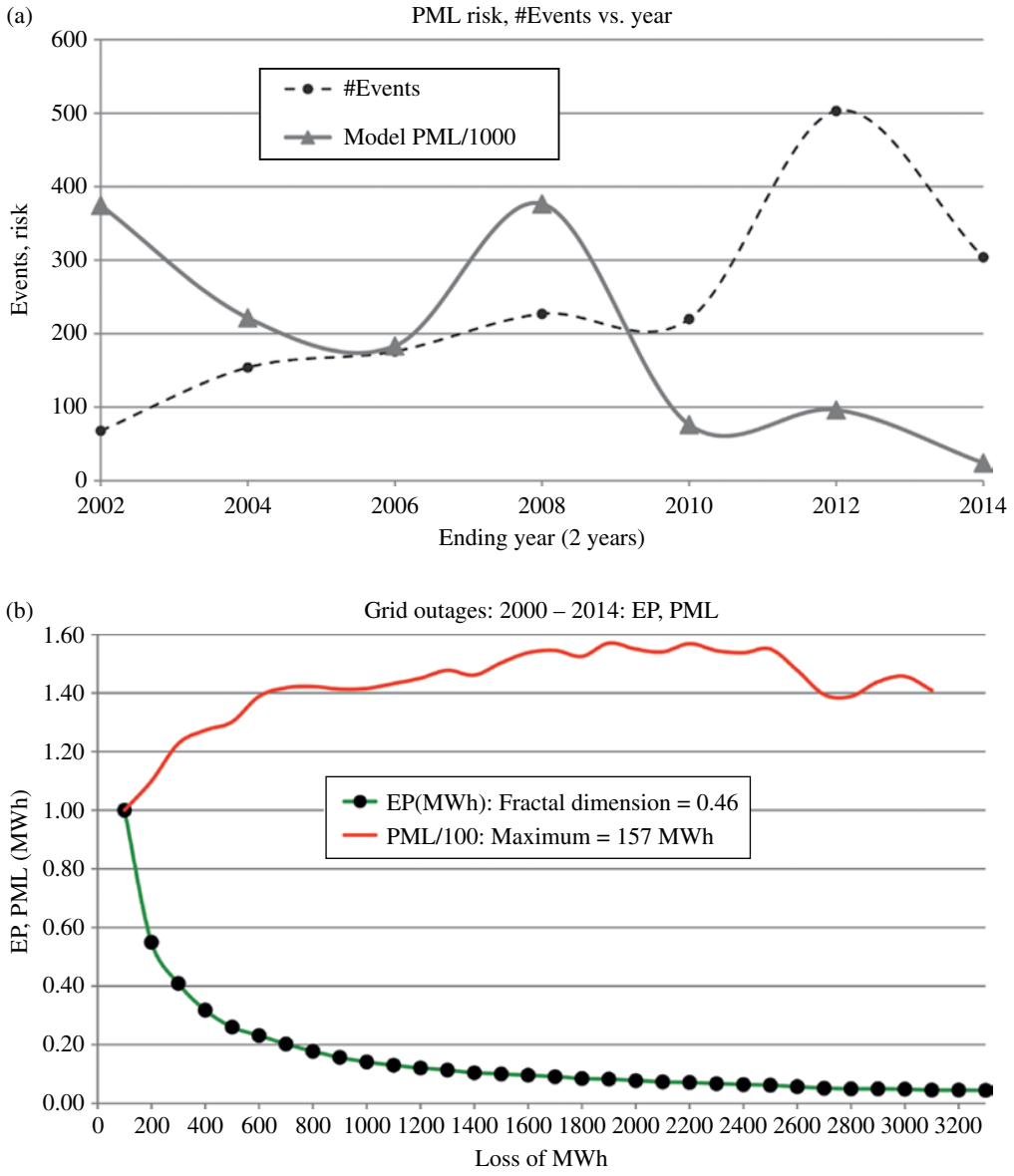


FIGURE 13.2 Power outages in the United States increased in number and size following deregulation but have declined since 2012. (a) Number of events peaked in 2012 and PML risk peaked in 2008. (b) PML risk is still high for the US power grid.

Eventually, the Tesla–Westinghouse approach won out and established AC as the standard technology for power generation and distribution. AC could be transmitted over longer distances than DC, easily powered motors used in factories and homes and could be voltage-stepped up/down to accommodate different needs for a diverse consumer.

By 1896 the Tesla–Westinghouse collaboration resulted in hydroelectric power generation at Niagara Falls and AC transmission to Buffalo 20 miles away. Edison's DC power networks were limited to 1 mile. This was the first Grid. It showed the technical and economic feasibility of electric power. Soon, privately owned and operated "power companies" sprang up across the nation. These companies were vertically integrated as shown in Figure 13.3. But over the course of a century, these vertically integrated power companies would be broken up into nonvertical oligopolies. The business of power distribution would take another 100 years to perfect.

13.2.1 Early Regulation

The first modern governmental regulator of all things having to do with energy and power—the *Federal Power Commission* (FPC)—was set up by Congress to coordinate hydroelectric projects in 1920. FPC grew over the decades and eventually become Federal Energy Regulatory Commission (FERC),

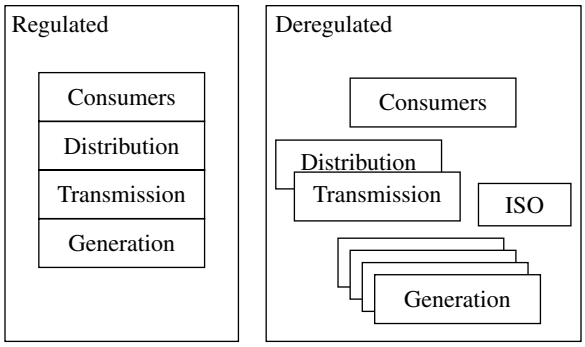


FIGURE 13.3 Vertically integrated power companies have been broken into oligopolies over the past 100 years.

with a budget exceeding $200 million and vast regulatory powers over natural gas and electrical power. But in the 1920s electrical power generation, transmission, and distribution was owned by large interstate holding companies that optimized the flow of power from fuels such as coal or hydroelectric generators. They exercised control of their regions of the country by vertically integrating all aspects of production, distribution, and marketing. See the regulated model of Figure 13.3.

The vertical monopolies standardized on 60 Hz (cycles/s) and 240/120-V current, but they were stovepiped islands when it came to interoperability. Two AC signals have to be synchronized before they can be combined across vertical monopolies. Synchronization would remain a technical challenge into the twenty-first century, including problems integrating solar and wind power into the Grid.

Standardization and synchronization was needed before privately held vertical monopolies could interoperate. *Universal access*—the ability for anyone in the United States to get electrical power service—had not yet arrived. It would require interoperability, a technical capability that was lacking among the local monopolies.

The Federal Power Act of 1920, the Natural Gas Act of 1938, and the Public Utility Holding Company Act (PUHCA) of 1935 changed the landscape by empowering the FPC to regulate the sale and transportation of natural gas and electricity across state borders. Together, these laws defined power and energy transmission as interstate commerce, which is the exclusive purview of the legislative branch of government. Thus a state could not directly regulate that commerce—but Congress could. PUHCA shaped the electric power industry until 1992.

A series of legal modifications to PUHCA expanded the power of Congress to regulate power and energy companies. For example, the Natural Gas Act was amended in 1940 to charge the FPC with responsibility for certifying and regulating natural gas facilities—going beyond simply regulating the sale of power across interstate boundaries.

The Northeast Blackout of 1965 highlighted the vulnerability of the vertically integrated power grid. As local holding companies were encouraged to interoperate and borrow power from one another to accommodate surges in demand, they also became more fragile. A loss of capacity in one region could lead to a series of failures that could collapse entire regions. Thus the cascade failure was born. Significantly, it forced a shift in federal regulatory legislation from pure regulation and universal access to an emphasis on safety and reliability.

The first prerequisite for prevention of cascade failures is that the power grid must be extremely reliable. Even a relatively insignificant component such as a power line must not fail. Thus the North American Electric Reliability Council (NERC) was formed shortly after the blackout in 1965. NERC is a not-for-profit company formed to promote the reliability of bulk electric systems that serve North America.[3]

The energy crisis of the 1970s brought fuel price inflation, conservation, and a growing concern for the environment. Congress began to shift its emphasis once again from reliability to clean and inexpensive power. The Public Utilities Regulatory Policies Act (PURPA) was enacted in 1978 to promote conservation of energy. But it had an important side effect: it opened the vertically integrated monopolies to competitors. PURPA required the electric utilities to buy power from "qualified facilities" (QFs). Thus was born the non-utility generator (NUG) and independent power producer (IPP). This side effect would be expanded in 1992 when the vertical monopolies were broken up by deregulation legislation.

In 1977, Congress transferred the powers of FPC into FERC—an independent agency that regulates the interstate transmission of natural gas, oil, and electricity. FERC maintained the shape of the electrical power sector during the 1980s and early 1990s. (For details on FERC's regulatory powers, see Chapter 12.)

FERC interprets and implements regulatory statutes that grant an exclusive franchise to electric utilities in exchange for low-cost *universal access* by all consumers. Universal access means that a lone farmer in a relatively sparse part of the country has access to electric power at the same cost as a city dweller surrounded by thousands of ratepayers. The cost of providing universal access was amortized over all consumers. This forced the monopolies into a "cost plus" business model rather than a model that encouraged innovation and expansion of power options. Universal

[3]"NERC's members are the 10 Regional Reliability Councils whose members come from all segments of the electric industry: investor-owned utilities; federal power agencies; rural electric cooperatives; state, municipal, and provincial utilities; independent power producers; power marketers; and end-use customers. These entities account for virtually all the electricity supplied in the United States, Canada, and a portion of Baja California Norte, Mexico," http://www.nerc.com.

access and regulation produced highly efficient, reliable, environmentally sensitive power, at the expense of technological advancement.

13.2.2 Deregulation and EPACT 1992

The era of regulated, layered vertical monopolies shown as in the regulated model of Figure 13.3 came to an end in 1992 with the enactment of the Energy Policy Act (EPACT). EPACT dramatically changed the industry once again. In addition to retaining clean, environmentally safe, reliable power, Congress now required utilities to provide "nondiscriminatory" transmission access to the transmission and distribution layers as shown in the deregulation model of Figure 13.3. Deregulation essentially replaced monopolies by oligopolies.

Under PURPA 1978, any QFs can use any part of the power grid to deliver its power to consumers. Under EPACT 1992, utilities are required to divest their interest in generation and open their transmission networks to any competitor. The intention paralleled other deregulations such as the 1996 Telecommunications Act, which promoted innovation by creating competition. Unfortunately, EPACT 1996 plunged the grid into chaos. According to one industry expert, the modern deregulated power industry is like a gasoline industry that fixes the price of oil at $30/barrel but allows the retail price of gasoline to go to $450/gallon!

> A particularly extreme example of the new sensitivity of prices occurred during the latter part of June 1998. For several days, spot-market prices for electricity in the Midwest experienced almost unheard-of volatility, soaring from typical values of about $25 per megawatt-hour (2.5 cents per kilowatt-hour) up to $7,500 per megawatt-hour ($7.50 per kilowatt-hour). Because the affected utilities were selling the power to their customers at fixed rates of less than 10 cents per kilowatt-hour, they lost a lot of money very quickly.
>
> The run-up in prices was so staggering that it might take an everyday analogy to appreciate it. In the 1970s, drivers howled when the price of gasoline tripled. Imagine your consternation if, one day, you pulled into a gas station and discovered the price had increased three hundredfold, from $1.50 per gallon to $450 per gallon.
>
> Most of us would look for alternative transportation. But with electricity you do not have options. With no way to store it, the affected utilities had a choice of either paying the going rate, or pulling the plug on their customers on the hottest day of the year. The total additional charges incurred by the utilities as a result of the price spike were estimated to be $500 million. [2]

As we shall see, this peculiar mixture of physics and economics will lead to vulnerabilities in the grid that must be considered when establishing policies for protection of this very critical infrastructure. In particular, the grid has been made more vulnerable at the point in history when it should be made less vulnerable. Economics has been given precedence over security. Deregulation encourages competition, but it discourages investment in the grid itself. The Grid currently suffers from the *tragedy of the commons*—a phenomenon described in greater detail in Chapter 3.

13.2.3 Energy Sector ISAC

The Electricity Sector ISAC should not be confused with EISAC—the Energy ISAC that deals with oil and natural gas information sharing.[4] ES-ISAC is run by the NERC and serves the electricity sector. It provides sharing among its electric sector members, federal government, and other critical infrastructure industries. Specifically, the mission of ES-ISAC is to collect and analyze security data and disseminate its analysis and warnings to its members, the FBI, and the Department of Homeland Security (DHS).

According to the ES-ISAC Web site:

> The Electricity Sector Information Sharing and Analysis Center (ES-ISAC) establishes situational awareness, incident management, coordination and communication capabilities within the electricity sector through timely, reliable and secure information exchange. The ES-ISAC, in collaboration with the Department of Energy and the Electricity Sector Coordinating Council (ESCC), serves as the primary security communications channel for the electricity sector and enhances the ability of the sector to prepare for and respond to cyber and physical threats, vulnerabilities and incidents.
>
> The ES-ISAC engages in the following activities:
>
> - Identifies, prioritizes and coordinates the protection of critical power services, infrastructure service and key resources
> - Facilitates sharing of information pertaining to physical and cyber threats, vulnerabilities, incidents, potential protective measures and practices
> - Provides rapid response through the ability to effectively contact and coordinate with member companies as required
> - Provides and shares campaign analysis, which includes capturing, correlating and trending data for historical analysis, and sharing that information within the sector
> - Receives incident data from private and public entities
> - Assists the Department of Energy, the Federal Energy Regulatory Commission and the Department of Homeland Security in analyzing event data to determine threat, vulnerabilities, trends and impacts for the sector, as well as interdependencies with other critical infrastructures (This includes integration into DHS' National Cyber security and Communications Integration Center.)
> - Analyzes incident data and prepares reports based on subject matter expertise in security and the bulk power system
> - Shares threat alerts, warnings, advisories, notices and vulnerability assessments with the industry

[4] http://www.esisac.com/

- Works with other ISACs to share information and provide assistance during actual or potential sector disruptions whether caused by intentional, accidental or natural events
- Develops and maintains an awareness of private and government infrastructure interdependencies
- Provides an electronic, secure capability for the ES-ISAC participants to exchange and share information on all threats to defend critical infrastructure
- Participates in government critical infrastructure exercises
- Conducts outreach to educate and inform the electricity sector[5]

13.3 OUT OF ORDERS 888 AND 889 COMES CHAOS

The EPACT of 1992 opened up the formerly closed transmission and distribution grid to all comers (FERC Order 889). The power companies of the vertically integrated era are now required to buy power from QF and allow competitors to use their transmission and distribution lines. But they can only charge consumers a usage fee set by state regulators—not them. Retail prices are fixed, while wholesale prices are allowed to float. The new grid is a competitive marketplace—almost. Floating wholesale prices can be inflated to the advantage of the seller, but each state sets retail prices as low as possible for political reasons. This has created chaotic economic shockwaves in states like California where power brokers have been allowed to "game the system" through predatory pricing contracts. Enron was perhaps the most notorious example of this practice.

The modern deregulated grid is still regulated for the purpose of encouraging innovation through competition. Still, it is a regulated industry with layers of regulators as shown in Figure 13.4. By Order 888, FERC created Independent System Operators (ISOs) that essentially replaced the monopolistic utilities with nonprofit "broker" companies. ISOs are where buyers meet sellers. According to Overby [2]:

> In a bid to ensure open and fair access by all to the transmission system, in Order 888 FERC envisioned the establishment of several region wide entities known as ISOs, or Independent System Operators. The purpose of the ISO is to replace the local utility's operation of the grid by a private, not-for-profit organization with no financial interest in the economic performance of any market players. In short, the job of the ISO is to keep the lights on, staying independent of and therefore impartial to the market players. As of the end of 1999 ISOs were operating the electrical grid in California, New England, New York, Texas and the coordinated power market known as PJM (Pennsylvania–New Jersey–Maryland).

[5]http://www.esisac.com/SitePages/Home.aspx

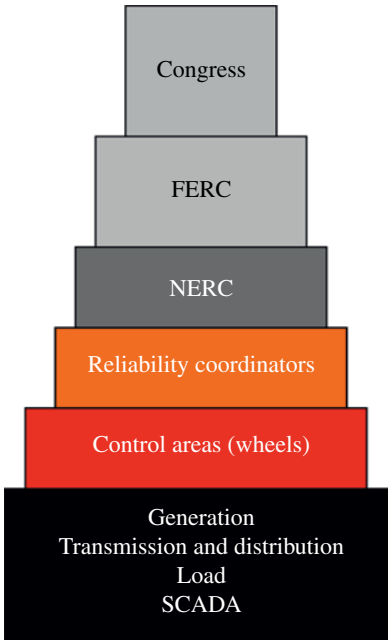


FIGURE 13.4 Many layers of regulation shape the Grid: Congress, FERC, NERC, Reliability Coordinators, Control Areas, and finally the laws of physics.

Under EPACT 1992, the responsibilities of an ISO are to:

- Control the transmission system.
- Maintain system reliability.
- Provide ancillary services such as system and voltage control, regulation, spinning reserve, supplemental operating reserve, and energy imbalance.
- Administer transmission tariff.
- Manage transmission constraints.
- Provide transmission system information (OASIS).
- Operate a power exchange (optional).

Sometimes the ISO separates the buying and selling activity from the regulation and reliability activities. In this case, they set up a separate power exchange. These are trading centers where utilities and other electricity suppliers submit price and quantity bids to buy and sell energy or services. Enron Online was one such exchange. It bought power on contract and resold its contracts to utility companies like PG&E in California. At one time, Enron Online cornered enough of the California market that it could charge whatever it wanted. This led to the California energy crisis in the late 1990s, which in turn led to the downfall of California's governor.

FERC requires an ISO to monitor its energy market for manipulation or abuses by the participants. This requirement covers both the power exchange (auction-based) market and bilateral transactions in the region (*wheeling*). An ISO's authority to take corrective action when market abuses are

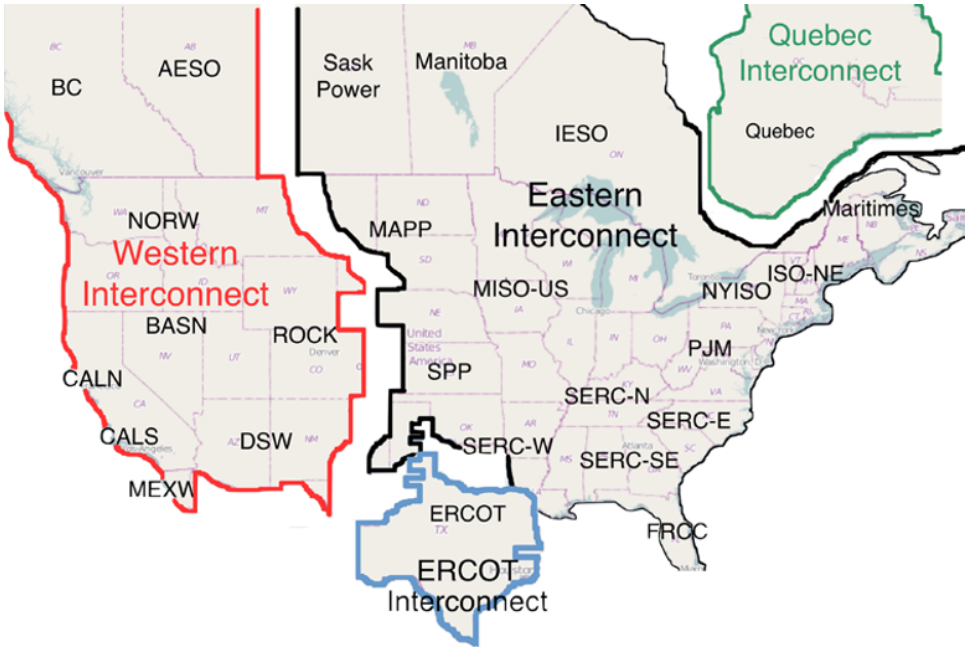


FIGURE 13.5 Major power grid interconnect components, reliability coordinators, and assessment areas of NERC include Canada and portions of Mexico [3]. Source: North American Electric Reliability Corporation.

identified depends on the nature of the abuse. In the case of abuses by Enron in 2002, the Department of Justice—not the ISO—pursued malfeasance charges against Enron executives.

Congress legislates and FERC regulates through cooperation with NERC. NERC has divided the United States and Canada into geographical areas called *Reliability Coordinators*. Each reliability coordinator oversees the operation of a number of *reliability assessment areas* sometimes called "wheels." Reliability coordinators monitor and adjust the flow of electrons throughout their region of responsibility. Buying and selling across control areas is called "wheeling" in the terminology of grid operators. The major reliability coordinators and control areas of North America are shown in Figure 13.5. Alphabetically, they are as follows.

NERC Reliability Assessment Areas

BASN: Basin (WECC)
CALN: California, North (WECC)
CALS: California, South (WECC)
DSW: Desert Southwest (WECC)
ERCOT: Electric Reliability Council of Texas (TRE)
FRCC: Florida Reliability Coordinating Council
ISO-NE: ISO New England, Inc. (NPCC)
MAPP: Mid-Continent Area Power Pool
MISO: Midwest Independent Transmission System Operator, Inc.
NORW: Northwest (WECC)
NYISO: New York Independent System Operator (NPCC)
PJM: PJM Interconnection
ROCK: Rockies (WECC)
SERC-E: SERC, East
SERC-N: SERC, North
SERC-SE: SERC, Southeast
SERC-W: SERC, West
SPP: Southwest Power Pool Regional Entity

Power is moved back and forth across these areas to balance supply and demand, but this balancing act is not always easy to do, because economics and physics do not always cooperate with one another. A surplus of electrons in one area may occur because of low demand, weather conditions, faults in transmission, or an overflow from another area. Operators have very limited options—they must either sell the surplus to an adjacent area or shut down generation. These options are not easy to achieve in a timely manner. Hence, physics often gets in the way of economics.

13.3.1 Economics Versus Physics

The economics of the deregulated grid often conflict with the laws of physics because:

- Electrons cannot be easily stored or inventoried—hence spot markets can be volatile, thus encouraging gaming of the system.
- The grid cannot easily redirect power to where it is needed—this foils demand and supply economics with both short-term and long-term implications.

- It is difficult to quantify the exact amount of power available at any point in time, which introduces human errors in the process of stabilizing the grid.
- A certain portion of the gird is "down" at any point in time because of maintenance, which makes it difficult for operators to estimate transmission and distribution capacity.

Economics and physics further clashed with politics as the Grid was deregulated throughout the 1990s and 2000s. A subtle SOC began building as a consequence—the physical distance separating power generator and customer began to increase. Power plants and solar farms were pushed away from populations to satisfy NIMBY, which required more transmission lines. It is politically easier to obtain permission to put solar farms in the desert, but politicians ignored the physics of transmission. NIMBY increases the load on transmission lines because remote power generation requires more transmission capacity. Thus, EPACT 1992 increased the load on an already overtaxed transmission network. Overloaded lines tend to burn out sooner, especially during the warm summer months. Taken together, NIMBY and EPACT have steadily increased SOC.

A subtle economic SOC also began to take over: *reactive power* began to go away, because power companies could no longer make money from it, and yet smooth operation of the grid depends on it. Reactive power is a form of electrical energy that sloshes back and forth between generator and load. Sloshing cancels the net–net transfer so there is no associated billing, hence there is no profit in maintaining reactive power. However, utilities must install heavier wires to handle the excess current—an added cost that saps profits. If power producers cannot profit from it, and utilities cannot charge for it, then why produce it? Without reactive power the grid became less stable.

A fourth conflict between economics, physics, and politics is emerging. Optimizing the grid by centralizing substations and power stations increases the network's spectral radius—another step toward the critical point. Distributing control among a handful of ISOs makes things even worse. In 2000 loss of grid capacity and control cost consumers $20 billion, but state public utility commissions refused to increase rates. Something had to give, so in the first few years leading up to the 2003 blackout, 150,000 skilled utility workers were let go. By August 14, 2003 the overextended operators of the Ohio portion of the Eastern Grid had inadequate situation awareness and inadequate options for handling a normal accident. As a consequence, a tripped line in Ohio toppled the Northern portion of the Eastern Grid, leaving 55 million people without electricity for more than 2 days. Airports, railroads, factories, hospitals, highways, Internet service providers, and emergency services were shut down across portions of northeastern Canada and the United States. At least 11 people died.

Note that Edison believed in distributed generation, which requires shorter transmission lines. Today's policies push in the wrong direction as they lead to more long-haul transmission—the opposite of Edison's design. As transmission lines become longer, the grid becomes less stable. Either we need to bolster long-haul transmission or return to Edison's original design. But then Edison never had to deal with NIMBY.

13.3.2 Betweenness Increases SOC

In 2005 a group of researchers constructed a network model of the US high-voltage transmission grid consisting of over 14,000 generators, transmission substations, and distribution substations and over 19,000 links representing transmission lines [4]. They then identified the betweener nodes—the nodes with the largest number of shortest paths passing to and from other nodes. These betweeners were the critical nodes of the grid. Out of the 14,000 nodes, only 140 were important enough to bring down the entire grid. Criticality in the power grid is highly correlated with node betweenness, and cascade fragility is highly correlated with large-scale outages. This is the clue we need to fully understand and analyze the Grid.

13.4 THE NORTH AMERICAN GRID

The North American Electric Grid is one of the largest and most complex man-made objects ever created. It consists of four large 60 Hz AC synchronous subsystems called the Eastern Interconnect, Western Interconnect (WSCC), Texas (ERCOT), and Quebec Interconnect. Figure 13.5 shows the four interconnects plus some of the subdivisions of each.[6]

The Eastern Interconnect has about 670,000 MW of capacity and a maximum demand of about 580,000 MW. The Western Interconnect has about 166,000 MW of capacity and a maximum demand of about 135,000 MW. ERCOT has 69,000 MW of capacity and maximum demand of 57,000 MW. Thus there is approximately 15% more generation capacity than demand at peak levels. The North American Electric Grid has sufficient power, but it lacks the transmission and distribution capacity needed to meet surge demand. This is a consequence of the historical development of vertical monopolies and the regulatory policies of Congress. It is also the grid's major weakness.

Theoretically the grid is able to move power from one place to another to meet demand. For example, peak power consumption in the Eastern Interconnect occurs 3 h before peak demand in the Western Interconnect simply because of time zones. In addition, weather conditions ameliorate the

[6]To see an animation of real-time flow of electricity in the Eastern Power Grid, visit: http://powerworld.com/Java/Eastern/

demand for power. During the winter Los Angeles sends power to heat homes in the Northwest, and during the summer Bonneville Power transmits power to Southern California to run air conditioners.

But this is theory. In reality, the grid is not robust enough to transmit power to where it is needed most. Instead, the grid has to be constantly monitored to meet demand and guard against cascading events such as tripped lines or power plants that are taken off line for maintenance. This challenge is mediated by SCADA/EMS at all levels throughout the grid.

The grid is made up of four major components: SCADA/EMS, generation, transmission and distribution, and consumer load. The last three are managed by various SCADA/EMS systems. Figure 13.1 illustrates this as a unified system of components called the grid:

1. *Generation*—source of electric energy: coal provides fuel for over half of the US electric power generators. There are more than 10,000 different generating units with a total capacity of about 800,000 MW in the United States. The largest generation plant is Grand Coulee Dam, Washington (7000 MW from hydro), and the next largest are Polo Verde, AZ (3700 MW from nuclear), W.A. Parish, TX (3600 MW of coal), and Scherer, GA (3400 MW from coal). Generation is fueled 56% by coal, 21% by nuclear, 9.6% by natural gas, 9.5% by hydroelectric, and 3.4% by petroleum. Most hydroelectric generators are in the East and West, most nuclear generators are in the Midwest and East, and thermal electric generation plants are spread throughout the United States.
2. *Transmission and Distribution*—the substations, transformers, and wires that carry the power from generation to load. There are more than 150,000 miles of high-voltage transmission lines in the United States. High-voltage lines operate at voltages up to 765 kV (kilovolt), with many 500, 345, and 230 kV lines. Higher-voltage lines typically consist of three wires attached to poles and towers by large conspicuous insulators. They are easy to identify from a passing automobile, bus, or train. Generally, they are in the open and unprotected. When a transmission line becomes too hot or shorts, it is said to have "tripped." Perhaps the most common fault in the grid stems from tripped lines. Often a line is overloaded in an attempt to shift power to where it is needed. The line heats up, sags, and touches a tree or the ground. Contact causes the circuit to short into the ground, and the line has to be shut down. Thus, a series of cascade failures can begin with a tripped high-power line. The greatest vulnerabilities of the grid are in the middle of the grid—its transmission and distribution network. The state of the transmission and distribution network is maintained by regional ISOs and the *Open Access Same-Time Information System* (OASIS) database. OASIS is an Internet-based database used by transmission providers and customers. It provides capacity reservation information, transmission prices, and ancillary services.
3. *Power lines have varying capacities*. The higher the voltage, the more efficient it is to transmit power. So generators deliver power to large-capacity, long-haul transmission lines (e.g., 733,000 V), which in turn deliver power to substations. The substations step the voltage down, say, to 230,000 V, and then transmit to other substations, which do the same. Finally, when the electricity reaches your back yard, it is reduced to 240/120 V. This is the idea behind the grid—use high-voltage lines to move power over long distances and low-voltage lines to move power around the consumer's home, factory, and so on.
4. *Load*—consumers are in complete control of demand; utilities must supply enough power to meet the load *at all times*. Total peak demand is about 710,000 MW, but the peaks occur at different times in different regions. In addition, demand can make dramatic swings—from 20,000 to 35,000 MW over a 1 week period and as much as 8,000–20,000 MW on an hourly basis. This means the SCADA/EMS system must be highly responsive and the operators must be alert. Gas-fired peaker plants are commonly used to meet surges in demand, but it may not be possible to distribute the additional capacity to where it is needed, because of inadequate transmission and distribution capacity. Hence, there is no shortage of power—but there is a shortage of transmission and distribution capacity. This, and the wild swings in demand, is the major reason for blackouts.
5. *SCADA and other control systems*—the control of all components of the grid. This component includes EMS and Power Plant Automation hardware and software. The main measure of how well the grid is doing is called the *area control error* (ACE). It is the difference between the actual flow of electricity into an area and the scheduled flow. Ideally, ACE should always be zero, but due to changing load conditions, ACE varies. Most wheels use *automatic generation control* (AGC) to adjust ACE. The goal of AGC is to keep ACE close to zero. Loss of a generator, transmission line, tower, or transformer can cause abrupt changes in ACE. It can take many minutes for AGC to rectify the loss and bring ACE back to zero. This is done by a complicated series of steps involving simulation of the intended change (say, to increase the power from a generator or buy power from an adjacent qualified facility). Power control systems work much like other sector's SCADA systems. Many remote

terminal units (RTUs) in the field collect data and control switches. The RTU data goes into a database, where EMS software calculates the next setting of the switches. And like other control systems, the control network sometimes hangs from the same towers and poles as the power lines themselves.

13.4.1 ACE and Kirchhoff's Law

The Grid is most vulnerable in the middle, the transmission and distribution layers, because there is insufficient capacity to deliver all the available power generated by the major interconnects. But more importantly, the Grid is a complex CIKR system subject to complex interactions due to physics. Kirchhoff's law says that at every point in an electrical circuit, the amount of electricity flowing in must equal the amount flowing out. Kirchhoff's law is another way of stating, "ACE must equal zero." System operators must chase ACE to meet unpredictable demand. This means either producing more power near the load or buying power from other parts of the grid. Under deregulation, they are encouraged to buy and sell from each other to drive ACE to zero. They are also required to allow competitors to use the old vertical monopoly's transmission and distribution layer. Add Kirchhoff's law to this dynamic balancing act and you risk destabilization of the Grid.

Consider the simple hypothetical grid shown in Figure 13.6 before and after a transmission line is dropped. Figure 13.6a models a simple generator-transmission-load-SCADA feedback network. In the lower right-hand corner is a node representing generators. At the top is a node representing the load, and in between are substations and transmission lines. One link from the load back to the generators represents a feedback signal that tells the generator operators to increase or decrease power in order to guarantee ACE = 0.

Every node in the network attempts to balance inflow with outflow. The inflows from all incoming links is summed and then apportioned equally to all outgoing links in honor of Kirchhoff's law. The network of Figure 13.6a will self-synchronize no matter what the inflows are. The network of Figure 13.6b will never synchronize regardless. Why?

Note that the link between Calvert and Minor (the middle link) has been removed in Figure 13.6b to simulate a dropped transmission line. This destabilizes the network so that it is impossible to obey Kirchhoff's law. Instead, electrical current oscillates forever as it flows through the damaged network. The reason is that Figure 13.6a contains an *aperiodic network*, while Figure 13.6b contains a *periodic network*. An aperiodic network self-synchronizes, while a periodic one does not.

An aperiodic network contains cycles of length m and n, where m and n are *relatively prime*. A *cycle* is a path from one node to others that returns to the starting node. Two integers, m and n are relatively prime if one divides the other with a remainder. For example, $m = 4$, $n = 3$ are relatively prime, because 4/3 = 1 with a remainder of 1.

There are five cycles in Figure 13.6a starting from and returning to the generator node:

1. Generator → Posum → SS13 → Red Bluff → Load → Generator: 5 hops
2. Generator → Calvert → SS5 → Minor → Load → Generator: 5 hops
3. Generator → Calvert → Minor → Load → Generator: 4 hops
4. Generator → Calvert → SS4 → Minor → Load → Generator: 5 hops
5. Generator → Calvert → SS4 → Annapolis → Load → Generator: 5 hops

Therefore, the Kirchhoff network in Figure 13.6a is aperiodic, because $m = 5$, $n = 4$ are relatively prime. But the Kirchhoff network of Figure 13.6b is periodic, because removal of cycle #3 leaves four cycles of length 5 hops and $m = 5$, $n = 5$ are nonprime relative to each other. This means instability in Figure 13.6b may not die out.

> ***Krichhoff Stability****: A Kirchhoff network is stable if it is aperiodic. Departures from Kirchhoff's law eventually die out and the network self-synchronizes.*

This illustrates the subtle complexities inherent in even the simplest power grid. Operators at each node (substation, power plant) may inadvertently destabilize a Kirchhoff network by attempting to balance ACE. Only a global understanding of the network's topology can overcome this error.

13.5 ANATOMY OF A BLACKOUT

According to Massoud Amin, Grid outages—called brownouts and blackouts—occur for a number of reasons, but in hindsight they are typically *normal accidents*. They start out as relatively insignificant faults or errors, which spread like a contagion to other power lines, substations, and power plants. Consequences grow as the outage sweeps across part, or all, of the Grid. The 2003 Blackout is a classic example of a normal accident.

The infamous Eastern Interconnection blackout of August 14, 2003, cut off power to millions of Americans and Canadians in eight states and one province. Lasting 2 days, the blackout shed 12% of NERC capacity. The economic costs—based on loss of electricity sales to consumers—range from $7 to $10 billion. The insurance industry lost $3 billion. Compare this with the damages and cleanup costs of TMI-2 (Three Mile Island Nuclear

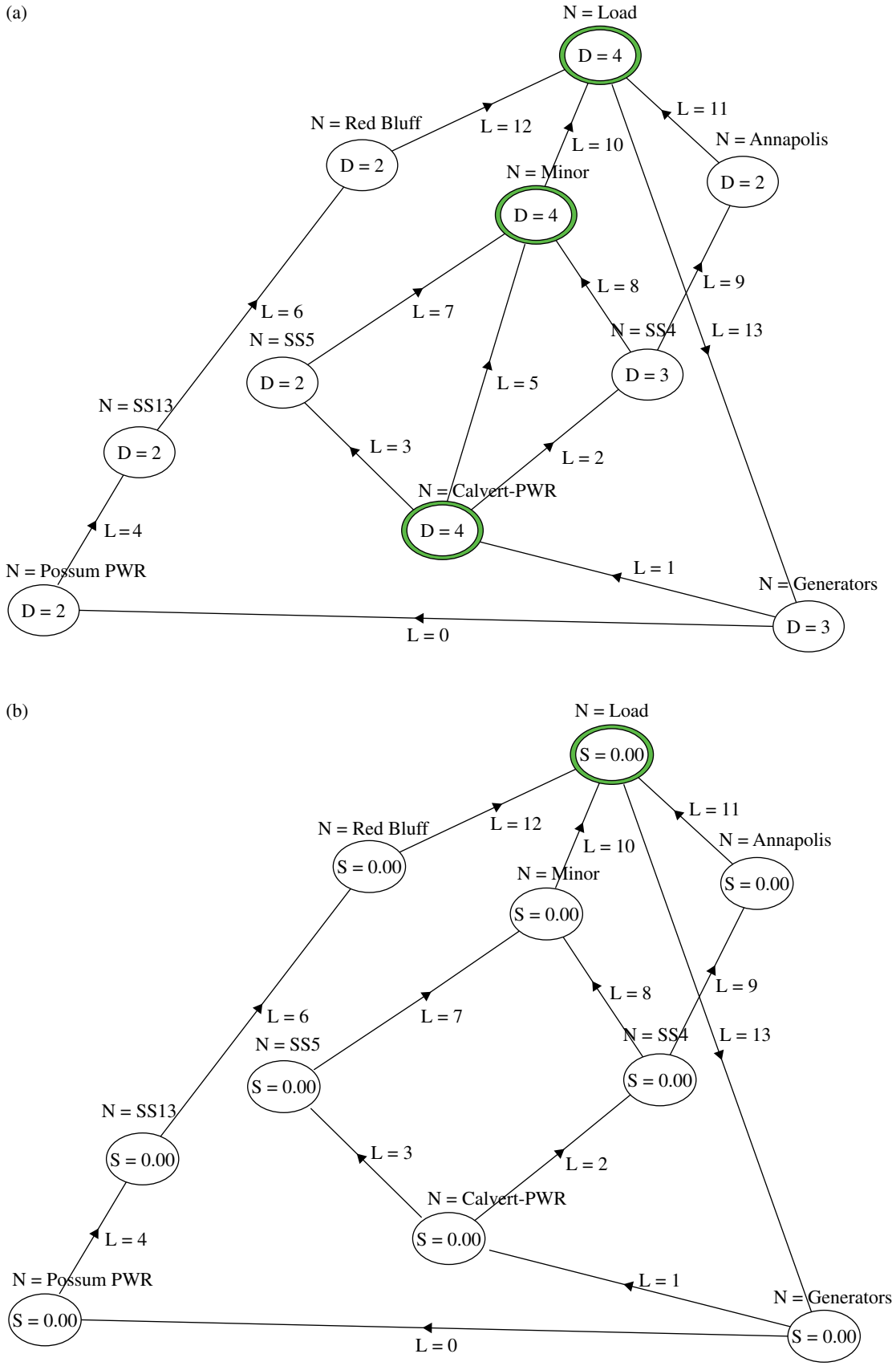


FIGURE 13.6 The Grid must obey Kirchhoff's law by adjusting inflows to equal outflows at all points in the network. A dropped link can destabilize the Grid so that it is impossible to make ACE equal to zero, as this hypothetical grid illustrates. (a) This hypothetical grid forms an *aperiodic network* that self-stabilizes. (b) This damaged grid forms a *periodic network* that is inherently unstable.

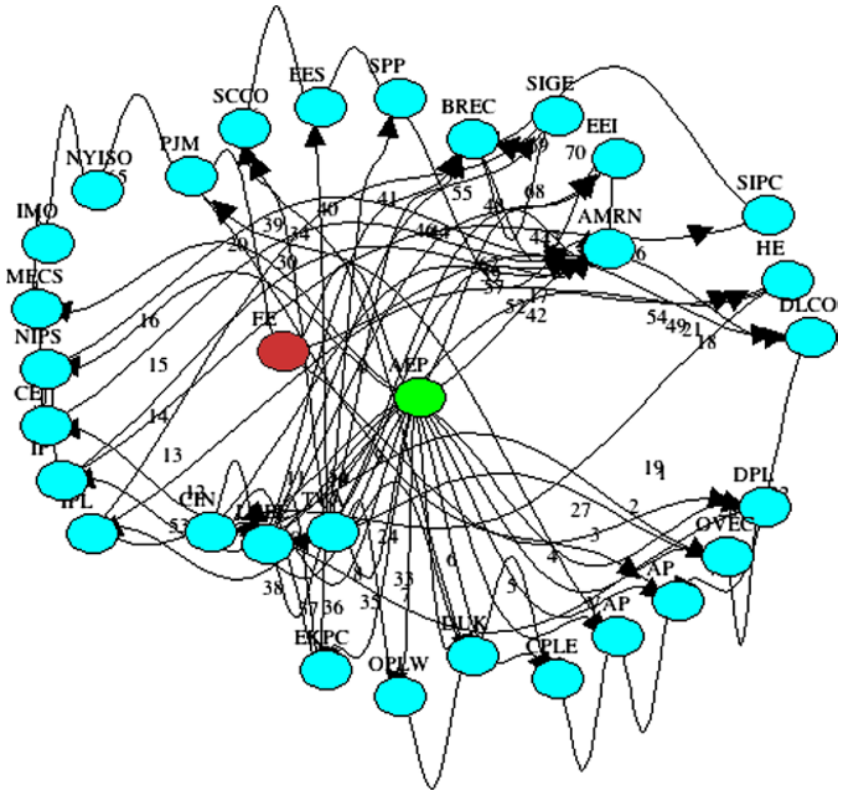


FIGURE 13.7 The 2003 Blackout started with reports from FirstEnergy (dark), which is part of the AEP (shaded) wheel at the epicenter of the blackout.

Power Plant #2) that melted down in 1979. TMI-2 incurred damages of $973 million—one-tenth the damage done by the Blackout of 2003.

According to the *Final Report of the US-Canada Power System Outage Task Force*, "The initiating events of the blackout involved two critical utilities—FirstEnergy (FE) and American Electric Power (AEP)—and their respective reliability coordinators, MISO and PJM." [5] (see Fig. 13.7). American Electric Power, Inc. (AEP) is an area within MISO (Midwest Independent System Operator). AEP of Columbus, Ohio, owns and operates more than 42,000 MW of generating capacity in the United States and in some international markets.[7] It is one of the largest electric utilities in the country, with almost 5 million customers linked to its 11-state electricity transmission and distribution network.

FirstEnergy Corp. of Akron, Ohio, is the fourth largest investor-owned electric power network in the United States. Its seven electric utility operating companies serve 4.3 million customers within 92,400 km^2 of Ohio, Pennsylvania, and New Jersey.[8] It also provides natural gas service to approximately 150,000 customers in the Midwest.

13.5.1 What Happened on August 14

The following sequence of events are broken down in stages so you can follow what happened and how cascade failures start out small and insignificant and grow, eventually overwhelming the entire grid.[9] The sequence of events leading to the outage is documented in greater detail in the *Interim* and *Final* reports produced by the US-Canada Power System Outage Task Force:

- FirstEnergy's control-room alarm system was not working, which meant operators did not know transmission lines had gone down, did not take any action to keep the problem from spreading, and did not alert anyone else in a timely fashion.
- MISO's tools for analyzing the system were also malfunctioning, and its reliability coordinators were using outdated data for monitoring—all of which kept MISO from noticing what was happening with FirstEnergy in time to avert the cascading.
- MISO and PJM Interconnection, the neighboring reliability coordinator, had no procedures to coordinate their reactions to transmission problems.

[7]http://www.aep.com
[8]http://www.firstenergycorp.com
[9]http://www.spectrum.ieee.org/WEBONLY/special/aug03/tline.html

The colossal collapse started out small—with an error in control software and tripping of an obscure power line in Ohio. Then the cascade unraveled in phases:

Phase I: Power Degradation

12:15 EDT: MISO SCADA/EMS state estimator software has high error—it is turned off.

13:31 EDT: Eastlake Unit #5 generation tripped in Ohio.

14:02 EDT: Stuart-Atlanta 345 kV line tripped in Ohio due to contact with a tree.

Phase II: Computer Failure

14:14: FE SCADA/EMS alarm software fails.

14:20: FE SCADA RTUs fail.

14:27: Star-South Canton 345 kV line tripped.

14:32: AEP called FE regarding Star-South Canton line.

14:41: FE transfers software applications to backup computer.

14:54: FE backup computer failed.

Phase III: Cascade Line Failures Begin

15:05: Harding-Chamberlin 345 kV line overheats, shorts with tree.

15:31: MISO called PJM to confirm Stuart-Atlanta line was out.

15:32: Hanna-Juniper 345 kV overheated, sags, and shorts out.

15:35: AEP unaware of Hanna-Juniper failure.

15:36: MISO unaware of Hanna-Juniper failure.

15:41: Star-South Canton tripped, closed, tripped again, unknown to AEP and PJM.

Phase IV: Cascading Collapses Transmission

15:39–15:58: Seven 138 kV lines trip.

15:59: Loss of the West Akron bus causes five more 138 kV lines to trip.

16:00–16:08: four more 138 kV lines trip; Sammis-Star 345 kV line overheats and trips.

Phases 5, 6, and 7

16:10–16:12: Transmission lines disconnect and form isolated islands in Northeast United States and Canada.

When it was all over, 263 of the 531 generators were shut down in the United States and Canada. The cascade that began in MISO spread to other regions: Quebec, Ontario, New England, New York, and Pennsylvania–New Jersey–Maryland (PJM). The 2003 Blackout was a normal accident that started with operator errors and tripped distribution lines and propagated to transmission lines and power plants. Eventually, 55 million people were without power.

This blackout qualified as a "1000 year flood," because of its size, measured along a number of metrics—it covered a large geographical area, affected a large population, and had a large economic impact. However, it was typical of many smaller-consequence outages that happen every day, because of:

- Power lines making contact with trees and shorting.
- Underestimation of generator output.
- Inability of operators to visualize the entire system.
- Failure to ensure operation within safe limits.
- Lack of coordination.
- Ineffective communication.
- Lack of "safety nets."
- Inadequate training of operators.

This handful of possible causes of blackouts ignores the potential for widespread and longer-term outages if the perpetrators are human. What if terrorists attempt to take down the Grid?

13.6 THREAT ANALYSIS

One way to identify human threats to the grid is to create "red team" scenarios by pretending to be a terrorist or criminal.[10] Maj. Warren Aronson, US Army, and Maj. Tom Arnold, US Marine Corps, prepared the following four attack scenarios while playing the role of red team.[11] They focused attention on power plant fuel supply, transmission line transformers, transmission substations including towers, SCADA, and power generators. These targets were chosen because they cost little to attack and yet they can create enormous damage or destabilize the Grid. If the red team can create an unstable grid, argued the red team, NERC rules require operators to propagate the instability across the entire control area and perhaps create a blackout across the entire interconnection.

13.6.1 Attack Scenario 1: Disruption of Fuel Supply to Power Plants

The process of supplying electricity begins with the transportation of power plant fuel by water, rail, road, or pipe to power generation plants. The largest source of North American electricity comes from coal-fired, thermal generating plants. These plants have historically maintained a reserve supply of 60–90 days of coal near each generator complex. Gas and fuel oil-fired plants generally have little, if any, fuel on-site, because of variance in seasonal demand for coal and dependence on just-in-time inventory to reduce

[10]Red teams are attackers, and blue teams are defenders.

[11]CS 3660 projects, summer 2002.

storage costs and environmental impact. This is an opportunity for an attacker.

A red team might disable, or at least significantly degrade, a major portion of regional power generation by attacking key components in the fuel supply chain. A specific example might be the Powder River Basin in Wyoming, described in the previous chapter. Only three railroad lines serve the region, carrying 305 million tons of coal annually to generation plants in more than a dozen states. Moving the same volume of coal by truck—currently the only alternative to rail—is both prohibitively expensive and restricted by available trucks and drivers, which currently support other consumers.

The destruction of an important bridge, like the *High Triple Bridge* over Antelope Creek, would stop coal transport on one of two primary lines feeding rail hubs for distribution to multiple states. Destruction of one to three similar targets immediately before peak periods of seasonal electricity demand could disable much of the country's generation capacity for periods of weeks to months. Vulnerability to this type of attack is high—say, 75%—and economic cost alone would exceed perhaps $2000 million.

13.6.2 Attack Scenario 2: Destruction of Major Transformers

Transformers are the key links between generation and transmission substations as well as between transmission and distribution subsystems. Most transformers are mechanically simple, consisting of wound copper coils encased in tanks of oil. The oil cools the coils to prevent the high voltage current from melting the copper, breaking the wire, and opening the transformer circuit. Step-up transformers servicing generation plants are very large and heavy, with some weighing hundreds of tons. Their size makes movement from the manufacturer to installation locations slow and expensive. Some transformers are made only in foreign countries and take months to replace. As a consequence, utility owners/operators and manufacturers do not maintain a large inventory. Step-down transformers can be equally difficult to replace and also represent choke points between transmission and distribution networks.

A devastating attack against step-up and step-down transformers is relatively simple. A single person can accomplish outright destruction quickly and inexpensively by planting explosives or driving a vehicle or material handling equipment into the side of a transformer. An even easier attack may be possible without entering the facility; puncturing the side of a transformer with a weapon like a rifle would cause coolant oil to leak, resulting in overheating before the attack is detected. Although heat sensors might shut down the transformer before fatal overheating, the loss of oil would temporarily stop the flow of electricity while the substation was shut down and transformer isolated and repaired. The consequences of a major transformer outage including economic consequences could exceed $100 million, and the probability of success is rather high—say, 95%.

13.6.3 Attack Scenario 3: Disruption of SCADA Communications

SCADA Operation Control Centers (OCCs) provide constant monitoring and adjustment to all subsystems of the electrical power system. Electric utility companies recognize the importance of these sites and have taken measures to protect them from physical attack. They are normally well protected and located in hardened structures, often behind layered security or below ground. Attackers could be insiders or militants that launch a direct assault on the facilities. However, recruiting existing employees sympathetic to the attacker's cause or placing a team member in a trusted position in such a facility requires total faith in that individual and may take considerable time. In addition, direct assault against a facility requires information on facility configuration, extended surveillance to discover security procedures, well-trained assault forces, overt action, and relative strength favoring the attacker, which is not typical of an asymmetric attack.

The weakest points in a control system are usually the communication networks themselves, rather than the OCC facilities. Although communication links normally have some form of redundancy, they are still susceptible to attack. Some components of the communication system will likely be exposed to observation and thus vulnerable to physical attack—for example, telephone wires strung on poles and externally mounted antennas. More sophisticated terrorists might use directed energy weapons or other forms of electronic warfare to damage SCADA without entering buildings.

Cyber exploits could target the published protocols used by SCADA systems, much like *Stuxnet* targeted the Siemens control system protocol. Regardless of the method chosen, the goal of these attacks is to both seize control of a power system and cause operations to occur outside safe operating parameters, destabilizing the ACE or disrupting recovery efforts following other events. The likelihood of this type of attack is comparatively low, say, 10%, but the consequences could be comparatively high, say, $1000 million.

13.6.4 Attack Scenario 4: Creation of a Cascading Transmission Failure

In accordance with NERC rules, generators and switching circuits are designed to automatically go offline when they operate outside safe operating ranges. Control circuits usually make automatic adjustments before the system exceeds these limits, but fail-safe devices will shut down generators and wheeling in the absence of external commands. When multiple failures occur nearly simultaneously, it is possible that the cumulative effect is an artificially induced normal

accident. The attacker's strategy is to induce a cascade by carefully chosen targets.

Here is how the normal accident might unfold. By design, if a major transmission line trips or substation fails, current surges and voltage transients trip circuit breakers. In reaction to the circuit breakers being tripped, the protective circuits at the affected generation plants shut down the turbines to prevent them from overspeeding due to the loss of load. As the load increases on the remaining generators, the rotation of the generator turbines decreases. As the turbine generators slow down, the power company must start to shut off some customers or bring more power online. If more power is not added or the load decreases to quickly, other generators will start to shutdown also. This cascade will continue until the entire grid is stopped.

The attacker must have enough expertise to know which substations and transmission lines to attack simultaneous. Therefore, the probability of success is low, say, 5%, and the consequences are uncertain, say, $500 million in equipment and economic loss.

13.7 RISK ANALYSIS

The forgoing red team scenarios are incorporated into the hypothetical grid of Figure 13.6a and input into the MBRA fault tree risk analysis tool as shown in Figure 13.8. For simplicity, assume all threats are 100% and all elimination costs are $100 million. Therefore, the only differences among threat–asset pairs are consequences and vulnerabilities as estimated by the red team.

Figure 13.9 shows the results of ROI analysis. Risk declines much faster than vulnerability, because of the OR-gate logic of the fault tree. An investment of $200 million nearly eliminates risk but lowers vulnerability to approximately 33%. The reason for the low vulnerability reduction ROI is traced to the high cost of protecting transformers. Transformer ROI is less than $1.00/$ invested. Therefore, transformer vulnerability remains high no matter how much of the limited budget is invested in protection of transformers.

13.8 ANALYSIS OF WECC96

The US power grid is vulnerable in the middle where transmission and distribution takes place. The importance of substations and transmission lines was underscored in 1996 when the Western Interconnect—also known as Western Electricity Coordinating Council (WECC)—was disrupted by a single failure in a transmission line connecting Oregon and California. This small fault spread throughout the 11 Western states pulling down the entire grid in a spectacular demonstration of normal accident theory.

Barabasi describes this spectacular failure in terms of network science:

> On a day of record temperatures, at 15:42:37 on August 10, 1996, the Allison-Keeler line in Oregon expanded and sagged close to a tree. There was a huge flash and the 1,300-megawatt line went dead. Because electricity cannot be stored, this enormous amount of power had to be suddenly shifted to neighboring lines. The shift took place automatically, funneling the current over to lower-voltage lines of 115 and 232 kilovolts, east of the Cascade Mountains. These power lines were not designed, however, to carry this excess power for an extended time. Loaded up to 115% of their thermal ratings, they too failed. A relay broke down in the 115-kilovolt line, and the excess current overheated the overloaded Ross-Lexington line, causing it too to drop into a tree. From this moment things could only keep deteriorating.

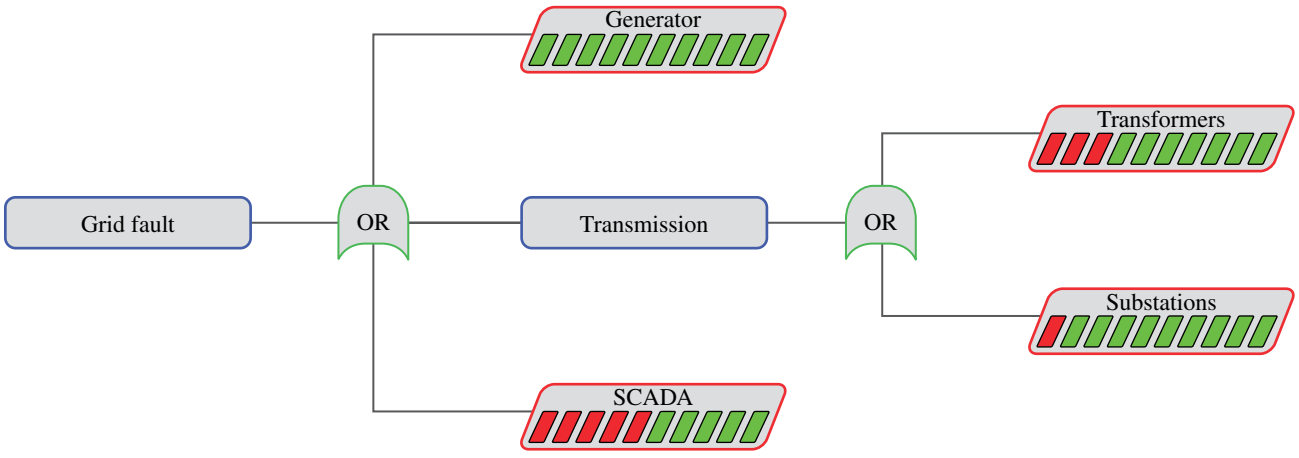


FIGURE 13.8 MBRA fault tree analysis of red team threats invests most in protecting scenario "Attack scenario 1: Disruption of fuel supply to power plants."

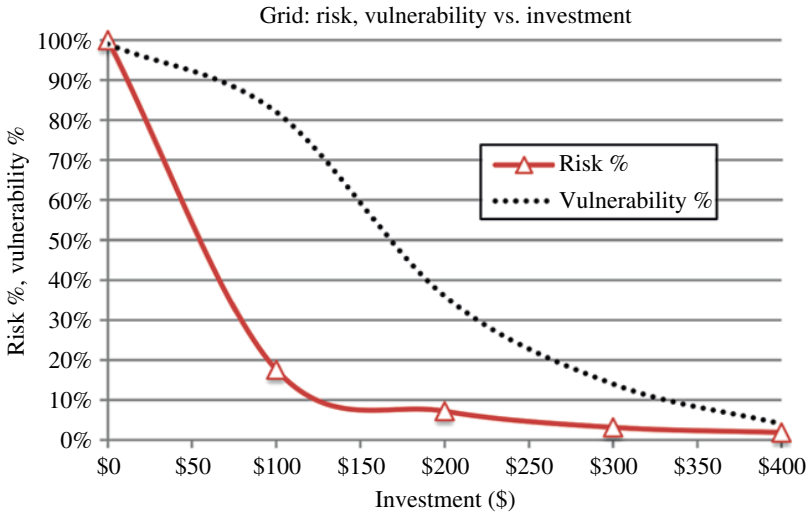


FIGURE 13.9 Threat analysis risk declines faster than vulnerability, because of the OR-gate logic of Figure 13.8.

> Thirteen generators at the McNary Dam malfunctioned, causing power and voltage oscillations, effectively separating the North-South Pacific Intertie near the California-Oregon border. This shattered the Western Interconnected Network into isolated pieces, creating a blackout in eleven U.S. states and two Canadian provinces. [6]

The 1996 power outage in the 11 Western states and Canada was prophetic. A similar failure on a relatively minor portion of the Eastern Grid led to massive outages in 2003. Knowledge of the 1996 outage did not prevent the 2003 blackout. This is because power engineers and politicians lack a thorough understanding of complexity theory as it applies to grid networks. Nonetheless, a number of complexity theory researchers identified the problem—complex CIKR network cascades are magnified by self-organized topology. In the case of the grid, self-organization is found in networks with high betweenness and high connectivity. The more connections a substation has, and the more paths passing through a substation or transformer, the more likely are disastrous cascade failures.

> ***Power Grid Resilience****: High values of spectral radius and betweenness in the network formed by substations, transmission lines, and interconnections decrease network resilience. To make the grid more resilient, spectral radius and betweenness must be minimized.*

The following analysis applies to WECC96—the WECC as it existed in 1996. This grid has been substantially improved since 1996. The numbers and analysis described here no longer applies, due to these improvements. However, it does explain why the Western power grid was unreliable prior to 2000.

The Western Interconnect as it was in 1996 shown in Figure 13.10 illustrates this principle. A network model of the WECC96 grid contains 181 nodes and 232 links, with node robustness of 22% and link robustness of 77%. Spectral radius is 3.46, which compared to mean connectivity of 2.56 is relatively mild. This network is comparatively resilient against cascades in theory. Over (0.22)(232) = 51 links have to be removed to separate the network into disjoint components, and (0.77)(181) = 139 nodes can be removed—one at a time—without separating the network into islands by single-node de-percolation. This means that (0.23)(181) = 41 blocking nodes form the critical nodes necessary to halt cascading or, alternatively, separate the network. Figure 13.11 shows the results of using the blocking node algorithm to identify which nodes to harden. Without these nodes, the Western power grid will not work, and with them, catastrophic cascade failures are possible because of their criticality.

The fundamental resiliency line, obtained by assuming random single node failures, indicates a high resilience to cascading through adjacent nodes. But flow resilience is low, largely because 52 of the 232 links are blocking links—removal of any one segments the network into islands.

WECC96 contained nodes and links with high combinations of connectivity (connectivity) and betweenness (4027 paths run through its hub). Connectivity measured by node connectivity promotes cascading, and congestion measured by betweenness promotes vulnerability. Together, these metrics signal potential weaknesses—hot spots—along paths between generator and load. The *hot spots*—areas in the network that are prone to failure because of high normalized connectivity times betweenness—are indicated with dark squares in Figure 13.10.

This analysis generally agrees with historical data indicating congestion zones, shown in Figure 13.10 as rectangular roadblocks. Square hot spots and rectangular congestion zones fall on major transmission paths running North and

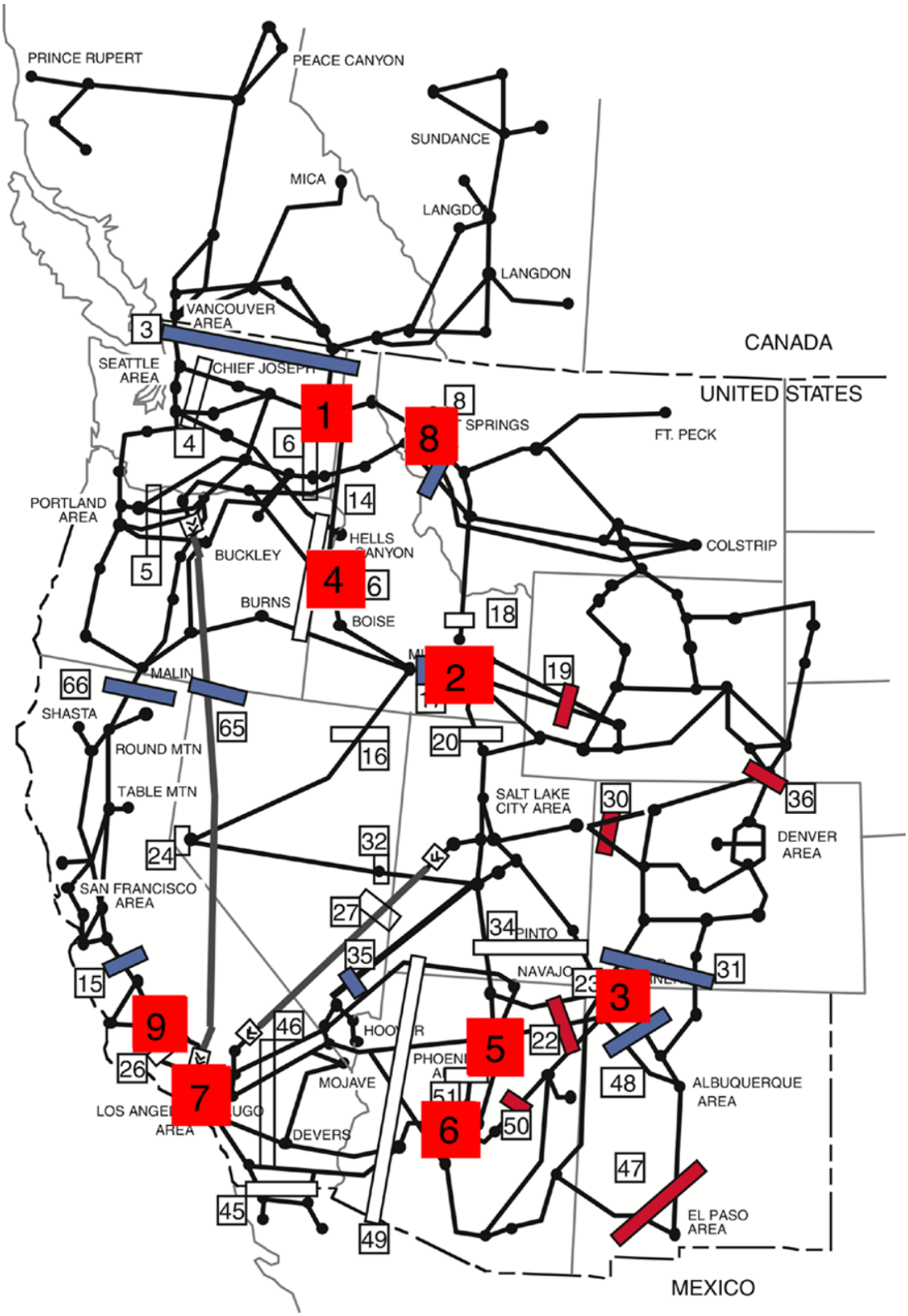


FIGURE 13.10 The Western power grid of 1996 (WECC96) contained a number of congestion zones as indicated by rectangular shapes and a number of high betweenness and connectivity hot spots as indicated by large (dark) numbered squares [7].

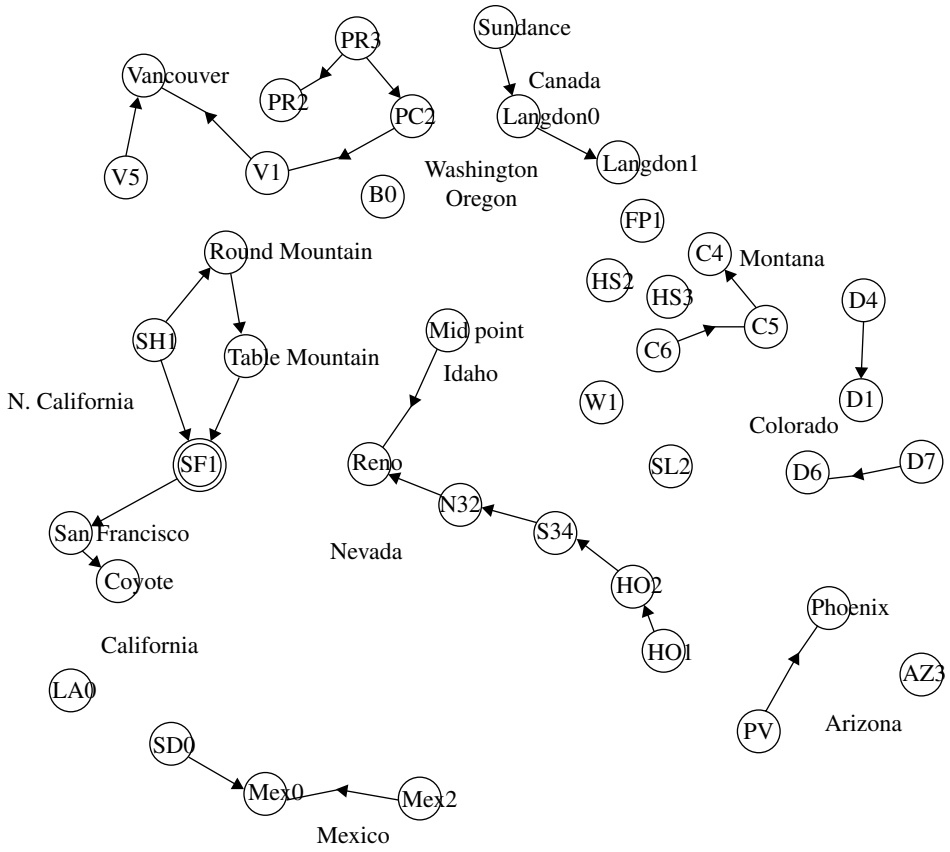


FIGURE 13.11 The WECC96 power grid's blocking nodes hold the grid together and also have the potential to stop cascade failures if hardened.

South between Arizona and Washington. During the warm season, power flows from the North to the South; during the cold months flow reverses—flowing from the South to the North. Los Angeles depends on Washington in the summer months, and Seattle depends on the Palo Verdes nuclear power plant in Arizona in the winter. Hot spots also lie on paths connecting power sources such as the dams at Bonneville to highly populated areas such as Seattle and Los Angeles.

The largest nuclear power plant is in Southern Arizona and the largest hydroelectric generator in the United States is located on the Columbia River separating Washington and Oregon. High-betweenness links connect these large power sources to large population centers—Phoenix, Denver, Los Angeles, Seattle, and points in between. If we want this grid to be more resilient, the hot spots in between generation and load must be eliminated. The way to do this is to rewire the WECC96 grid such that node connectivity and betweenness are reduced.

13.9 ANALYSIS

Theoretically, a large power grid can shift power from one end of the country to the other because it can be extremely adaptable to changing demand and localized faults. Even if the largest dozen or so centralized power plants fail, power can theoretically be transferred from somewhere else. The largest dozen power plants supply less than 5% of national power demand, and there is a 15% surplus at any moment in time. In addition, different regions of the grid reach peak load at different times, so when demand peaks in one part of the grid, the demand can be satisfied by a demand valley in another part.

So, the larger the grid, the more adaptable it is—theoretically. But in practice, the grid is too complex to guarantee isolation of faults (versus cascading), and vertical integration over the past century has led to regional interoperability problems. Simply put, it is still impossible for the grid to adapt to demand on a national scale because there is insufficient capacity in the transmission and distribution network. To make matters worse, the SCADA/EMS systems are not sophisticated enough to properly automate the regulation of ACE. While there is no shortage of power, there is a shortage of distribution capability, SCADA sophistication, and trained operators.

The 1992 EPACT was aimed at decoupling the layers of the old vertical monopolies, but at the present time, this has increased, rather than decreased, the vulnerability of

the grid. In addition to economic vulnerability (the "gaming of prices" by predator utility brokers like Enron), the network is vulnerable to technical vulnerabilities (SCADA software errors, complex interdependencies that are not fully understood). EPACT has deregulated the generation and load components but left the middle component on its own in a world that views the transmission and distribution component "someone else's problem." There is no money to be made from the middle. Thus, economic forces are working against protection of the most vulnerable part of the grid.

The grid may simply be too big and complex to fully control. In fact, the grid may not be entirely necessary. In 1902 there were 50,000 independent, isolated power-generating plants in the United States and only 3,624 central power plants [8]. The grid started out as a decentralized, distributed generation network. Immediately after World War I, the price of coal soared and urbanization favored centralization of generation. Technology advanced rapidly during the 1920s, which lowered the cost of building centralized plants. In addition, their owners drove the independents out of the market by lowering monthly bills to consumers and diversifying applications to make up for the loss of revenue during non-peak periods of the day. The vertically integrated and centralized power companies sold their non-peak power to electrified train systems (subways and commuter trains), factories, and large building owners to raise elevators in tall buildings. Thus, centralized generation won out, and today we have a grid with a high connectivity of SOC.

But the grid does not have to remain the way it is today. If it was redesigned and regulatory legislation was to favor *distributed generation* (wind, solar, and fuel cell generators at factories, shopping malls, neighborhoods), the grid would be made almost invincible because it would truly be adaptable. In distributed generation systems, most of the time most of the power comes from only a few yards away. Solar generators do not produce during the night, and wind power does not produce during periods of calm weather, so the grid might still be needed. But it would be needed less of the time, and when it fails, the local generation facility would provide enough power to keep critical services like hospitals operating.

In addition, large storage cells located close to metropolitan centers would further alleviate the burden on transmission lines. During off-peak periods, generators could use the transmission lines to charge up batteries, flywheels, or reservoirs. During peak demand periods, power could be drawn from local storage rather than power plants located hundreds of miles away. Inadequate transmission and distribution capacity would become less critical.

This leaves SCADA/EMS as the vulnerability of greatest concern. And the cyber threat to power is real. The SQL Slammer worm penetrated a private computer network at Ohio's Davis–Besse nuclear power plant in January 2003. It disabled a safety monitoring system for nearly 5 h and shut down a critical control network after moving from a corporate network, through a remote computer onto the local area network that was connected to the control center network. SQL Slammer could have affected critical control systems at Davis–Besse. As it turned out, the affected systems were used to monitor, not control, the reactor. The safety of Davis–Besse was not jeopardized.

By 2005, more than 60 cyber security events impacted power control systems, including three nuclear plants.[12] This number is likely to grow as the Internet becomes intertwined with non-Internet control networks. Unfortunately, SCADA/EMS components—computers, networks, and software—will remain complex and unreliable for a long time because securing an information system is well known to be problematic. Thus far, it has been impossible to build software that is guaranteed to be bug-free. These software flaws lead to networks becoming disconnected, data being lost, and computers being disabled. As long as software is flawed, there will be faults in industrial control systems such as SCADA and EMS. And, as long as software is designed and written by humans, it will be flawed.

13.10 EXERCISES

1. Why are there high- and low-voltage lines?
   a. Cities need more power than farms.
   b. Electrons travel farther on high voltage.
   c. Electricity travels more efficiently at high voltage.
   d. Electrons travel faster at high voltage.
   e. Electricity has lower resistance at low voltage.
2. AC won over DC because AC:
   a. Works better in radios and TVs
   b. Is an international standard
   c. Operates at 60 cycles/s, which is compatible with clocks
   d. Can be transmitted at high voltages
   e. Can be switched like the Internet
3. Before it was called FERC, it was called:
   a. FPC
   b. CIA
   c. NERC
   d. MISO
   e. NCS
4. The PUHCA of 1935 established federal regulatory control over power because:
   a. It was the right thing to do
   b. Rural areas needed power too
   c. The Great Depression was in full effect
   d. Congress wanted to establish power over power
   e. Interstate commerce allows the federal government to regulate sales

[12] A personal communication with Joe Weiss.

5. NERC and load sharing through wheeling was established soon after:
   a. Enactment of the Federal Power Act of 1920
   b. The Northeast Blackout of 1965
   c. Soon after the problem of synchronization was solved in the 1970s
   d. Enactment of PURPA in 1978
   e. Soon after deregulation in 1992
6. Electrical power was deregulated by enactment of:
   a. EPACT 1992
   b. PURPA in 1992
   c. Establishment of ISOs in 1992
   d. Tragedy of the Commons Act of 1992
   e. ISOs
7. ISOs were authorized by FERC to:
   a. Monitor the operators
   b. Look for abuses by participants
   c. Run power exchange markets
   d. Maintain their independence
   e. All of the above
8. The Electricity Sector ISAC (ES-ISAC) is:
   a. The same as EISAC
   b. Run by NERC
   c. Run by FERC
   d. Run by ISO
   e. Run by the Department of Homeland Security
9. Which one of the following is *NOT* a power grid within NERC?
   a. ERCOT
   b. Western Interconnect
   c. Quebec Interconnect
   d. Midwestern Interconnect
   e. Eastern Interconnect
10. Which of the following is *NOT* a component of the US electrical power grid?
    a. Dams
    b. Power plants
    c. Transmission
    d. Distribution networks
    e. SCADA
11. Which of the following make the electrical power grid particularly vulnerable?
    a. Most power comes from a few central power plants.
    b. Coal fuel supplies in the United States depend on critical railroad links.
    c. Large transformers never break.
    d. SmartGrid technology.
    e. Hydroelectric dams are vulnerable.
12. In the United States, power outages have been:
    a. Constant
    b. Increasing
    c. Decreasing
    d. Smaller than in other countries
    e. None of the above
13. The most asymmetric attack on power generation would be:
    a. Bombing of Grand Coulee Dam
    b. Attacking a nuclear power plant
    c. Coordinated attack on substations using fault trees
    d. Cyber attack on SCADA systems that control power
    e. Bombing of Hoover Dam on the Colorado River
14. Critical transmission paths are defined by:
    a. Towers carrying high voltage power
    b. Interstate tie lines
    c. Local distribution network transformers
    d. Transmission lines supplying power to major areas such as Chicago
    e. High connectivity and betweenness hot spots
15. Large transformers are considered critical, because they are:
    a. Difficult to transport from manufacturing to installation
    b. Easy to destroy
    c. Cause power outages
    d. Expensive
    e. All of the above
16. Why does ACE deviate from zero?
    a. The load is constantly changing.
    b. Generators generate unpredictable output.
    c. SCADA/EMS software often fails.
    d. The weather is constantly changing.
    e. The grid is too big and complicated to understand.
17. Why is the Grid vulnerable in the middle?
    a. There is insufficient transmission and distribution capacity.
    b. Transformers are critical and unprotected.
    c. Substations are unreliable.
    d. Everything depends on generators.
    e. Fuel is in short supply.
18. Deregulation under EPACT 1992 allowed:
    a. Utilities to make more money
    b. Competing utilities to use the transmission infrastructure
    c. Increased build-out of more transmission lines
    d. Reduced blackouts
    e. Higher penalties for letting ACE deviate from zero
19. A Kirchhoff grid can be destabilized if it forms a:
    a. Periodic network
    b. Aperiodic network
    c. Scale-free network
    d. Clustered network
    e. Separated network

20. Distributed generation solves the problem of:
    a. Renewable energy sources
    b. Underutilized transmission
    c. Transmission capacity
    d. NIMBY
    e. Regulation

13.11 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. The PML graph of Figure 13.2b shows a very long-tailed exceedence probability with low fractal dimension. Is this considered high or low risk? Explain your answer.
B. Renewable energy sources like wind and solar are not only cleaner, but wind and sun are free! So, why does the electric power sector—utilities, mostly—object to converting coal and gas-powered generators to windmills and solar panels?
C. If electric power is unstable because of a shortage of transmission lines, why not build more? Explain why more transmission lines are not being constructed in the United States.
D. Blocking nodes become useful for blocking the spread of a cascade failure, but are they also critical for propagating the flow of electrons? Explain why blocking nodes are critical for blocking and not blocking at the same time.
E. Freeways and interstate highways are free to use by anyone, but electrical power lines are not. Explain why private corporations own the highly important power transmission lines and the highways (for the most part) are not.

REFERENCES

[1] Lewis, T. G. *Bak's Sand Pile*, Monterey: AgilePress, 2011.
[2] Overbye, T. Reengineering the Electric Grid, *American Scientist*, 88, 3, May–June 2000, pp. 220.
[3] U.S.-Canada Power System Outage Task Force. Interim Report: Causes of the August 14th Blackout in the United States and Canada, November 2003, pp. 134.
[4] Kinney, R., Crucitti, P., Albert, R., and Latora, V. Modeling Cascading Failures in the North American Power Grid, *European Physical Journal B*, 46, 1, 2005, pp. 101–107.
[5] U.S.-Canada Power System Outage Task Force. Final Report on the August 14, 2003, Blackout in the United States and Canada: Causes and Recommendations, April 2004, pp. 12. https://reports.energy.gov (dated January 12, 2004).
[6] Barabási, A.-L. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*, New York: PLUME, 2003, pp. 119.
[7] US Department of Energy. National Electric Transmission Congestion Study, August 2006, pp. 32.
[8] Friedlander, A. *Power and Light: Electricity in the U.S. Energy Infrastructure 1870–1940*, Reston: Corporation for National Research Initiatives, 1996, pp. 51.

14

HEALTHCARE AND PUBLIC HEALTH

According to the Department of Homeland Security, "The Healthcare and Public Health Sector protects all sectors of the economy from hazards such as terrorism, infectious disease outbreaks, and natural disasters. Because the vast majority of the sector's assets are privately owned and operated, collaboration and information sharing between the public and private sectors is essential to increasing resilience of the nation's Healthcare and Public Health critical infrastructure. Operating in all U.S. states, territories, and tribal areas, the sector plays a significant role in response and recovery across all other sectors in the event of a natural or manmade disaster. The Healthcare and Public Health Sector is highly dependent on fellow sectors for continuity of operations and service delivery, including: Communications, Emergency Services, Energy, Food and Agriculture, Information Technology, Transportation Systems, and Water and Wastewater Systems."[1]

The Department of Homeland Security issued the Healthcare and Public Health Sector-Specific Plan (HPH SSP) in 2010. The introductory portion of this chapter summarizes this plan, and the remainder of this chapter emphasizes bioterrorism and epidemiology—the focus of homeland defense and security. The major results of this emphasis are as follows:

- *HPH is a complex CIKR:* Healthcare and public health (HPH) is a complex system spanning interdependent state, local, tribal, and federal government agencies and both public and private organizations. Additionally, it consumes 17% of the US economy (GDP) and aims to provide a broad array of services to citizens during periods of relative calm and during disasters. It interacts with transportation, communications, energy, water, emergency services, information technology, chemical, and food and agriculture sectors. It may be too ambitious and costly, because healthcare costs are rising faster than the general economy as measured by gross domestic product (GDP).
- *Goals of HPH sector:* The HPH sector is the largest industrial commons in the United States employing 13 million workers. It aims to provide supplies and services during emergencies, protect healthcare workers while on duty, and mitigate risks to physical and cyber assets, and in addition to providing non-emergency services such as vaccinations and data collection/dissemination, the sector operates in a complex socioeconomic and political environment subject to tragedy of the commons forces. Its roots began in 1938 with the Food, Drug, and Cosmetic Act and the 1944 Public Health Service Act and continue to evolve today as the Affordable Care Act unfolds.
- *Roemer's model:* Roemer's model is a device for understanding this complex CIKR. The HPH sector is composed of five components—management, organizations, resources, delivery of services, and economic support. One-third of all funding for public health services comes from the federal government in the form of Medicare and Medicaid. And yet, this sector is organized mainly around the private sector—the medical industrial commons made up of private practices, pharmaceuticals, and insurance companies.

[1]http://www.dhs.gov/healthcare-and-public-health-sector

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

- *HSPD-21:* Presidential directive HSPD-21 changed the direction of the HPH sector by emphasizing proactive programs in bio-surveillance, drug stockpiling, mass casualty preparedness, and community resilience. The sector became even more controversial in the United States as the Affordable Care Act began to change the focus from a mostly private sector commons to a public sector commons. Yet, sustainability of this sector remains in question as costs outstrip economic progress. Funding is the number one threat to this sector.
- *Bioterrorism:* Attacks on US soil and the threat of pandemics from abroad are driving the transformation of HPH according to HSPD-21 dictates. International air travel, concentration of people in ever-larger cities, and the rise of new strains of disease require a greater emphasis on bio-surveillance and countermeasure techniques.
- *Causes of death:* Data on causes of death and factors in the causes indicates that smoking and poor diet/exercise are by far larger threats to the HPH sector than bioterrorism, air travel, or global pandemics. Nearly 50% of all deaths reported in 2000 were due to heart disease and cancer. But smoking and poor diet/exercise—lifestyle choices—were responsible for one-third of the underlying factors. The public is mostly vulnerable to lifestyle choices, which result in costly deaths.
- *Models:* The historic Kermack–McKendrick [3] model of the spread of diseases through contact served epidemiologist well for over 80 years, but the rise of air travel, densely settled cities, and new strains of diseases renders the Kermack–McKendrick model inadequate. In its place is a collection of network-based models that represent vulnerable populations as social networks. Connectivity is what matters in these new models.
- *Blocking countermeasures:* Self-organization in the form of percolation, and structure in the form of blocking nodes, determines the rate and extent of spreading in a social network. A conservative countermeasure strategy hardens the n/ρ critical blocking nodes. A more ambitious strategy eliminates or protects redundant links by de-percolation. However, a practical strategy hardens the highest-ranking nodes as determined by normalized degree and betweenness.
- *Bio threats:* Biological threats are classified as bacteria, rickettsiae, virus, fungus, and toxins. The Center for Disease Control (CDC) categorizes pathological and hazardous agents according to their impact on public health. Category A agents are the most threatening, Category B is next, and Category C is least threatening. Examples of Category A agents are anthrax, smallpox, bubonic plague, and hemorrhagic fevers.
- *SARS:* A case study of severe acute respiratory syndrome (SARS) suggests that a quick response to a potential pandemic is the most effective strategy for halting the spread of a global contagion. On the theoretical side, research suggests that the spread of SARS was mitigated by air travel rather than the opposite. A group of researchers in China claim that long Levy flights by contagious diseases like SARS and H1N1 reduce virulence. SARS halted because it jumped too far, too fast, as it spread to 29 countries.
- *Air travel network:* The spread of contagions through the OpenFlight500 network, consisting of the top 500 airports as nodes and the 4096 routes as links, can be controlled by hardening blocking nodes, de-percolating routes, or hardening of the top 20 airports. Simulation of OpenFlight500 suggests that hardening the top 20 airports is the most practical countermeasure.
- *Pandemic factor:* Spectral radius and infectiousness determine the virulence of a potential pandemic. If $\gamma\rho > \Delta$, the disease will spread without bound. Infectiousness cannot be controlled, but spectral radius can. Therefore, strategies that reduce ρ also reduce spreading. Spectral radius can be reduced by de-percolation, removal of blocking nodes, and hardening of super-spreader nodes as indicated by normalized degree and betweenness. The most economical countermeasure is the latter—hardening of super-spreaders.

14.1 THE SECTOR PLAN

HPH is the largest industry in the United States. It consumed more than 17% of the national economy in 2010—an astronomical amount expected to rise because of complex interactions between and among social, economic, political, and technical factors. The rapid rise of a senior citizen class with longer life span and higher expectations of healthcare, the rising cost of healthcare, the shift from privately funded healthcare to government funded, and the exponential improvement in medical technology have all conspired to make this sector especially complex. Unlike most critical infrastructure, this sector is likely to undergo more rapid socio-technical change than any other sector.

The 2010 *Healthcare and Public Health Sector-Specific Plan* co-developed by the Department of Health and Human Services (DHHS) and the Department of Homeland Security has an ambitious goal of protecting the public against all hazards—defined as "natural disasters, pandemics, terrorist attacks, and other manmade disasters":

> The HPH Sector will achieve overall resilience against all hazards. It will prevent or minimize damage to, or destruction of, the Nation's healthcare and public health infrastructure. It will strive to protect its workforce and preserve its ability to

mount timely and effective responses (without disruption to services in unaffected areas) and to recover from both routine and emergency situations.[2]

It does this using a risk-informed decision-making strategy applied to the principle components of healthcare:

Supply and service protection—provide essential health services during and after disasters or disruptions in supplies or supporting services, for example, medicines, water, and power.
Workforce protection—protect healthcare workers from consequences that may compromise their health and safety and limit their ability to carry out their responsibilities.
Physical asset protection—mitigate the risks posed by all hazards to the sector's physical assets.
Cyber security protection—mitigate risks to the sector's cyber assets that may result in disruption to or denial of health services.

The HPH sector is a networked system consisting of 13 million healthcare workers, 4,000 hospitals, 500,000 ambulatory service organizations, 75,000 nursing facilities, 42,000 retail pharmacies, federal and state public health departments, 1,100 drug companies, 2,500 medical device manufacturers, 1,200 blood and organ banks, and thousands of health insurance companies. Most of this complex is owned and operated by the private sector but is heavily dependent on policies established by federal and state policies.

To complicate matters even more, this sector is highly interdependent with transportation, communications, energy, water, emergency services, information technology, chemical, and food and agriculture. Transportation is essential for movement of people and supplies, communications for support of emergency operations, energy to power hospitals and transportation, emergency services for coordination with EMS and law enforcement during catastrophic events, chemicals in support of pharmaceuticals, and food and water for human survival and healthcare. These sectors intersect and interoperate in complex ways. For example, power outages during hurricanes or floods can render hospitals useless and block the movement of first responders to where they are needed.

14.2 ROEMER'S MODEL

HPH has a long history going back to the passage of the Food, Drug, and Cosmetic Act of 1938 and Public Health Service Act of 1944. In modern times, public health policy has been shaped by bioterrorism and bioterrorism legislation. The Project BioShield Act of 2004 authorized $5 billion to purchase and stockpile vaccines that would likely be needed as a result of a terrorist attack. In the first decade of this act, over $50 billion was spent "to provide protections and countermeasures against chemical, radiological, or nuclear agents that may be used in a terrorist attack against the United States by giving the National Institutes of Health contracting flexibility, infrastructure improvements, and expediting the scientific peer review process, and streamlining the Food and Drug Administration approval process of countermeasures."[3] The *Pandemic and All-Hazards Preparedness Reauthorization Act of 2013* renewed funding of BioShield. Reportedly, the DHHS has large stockpiles of antitoxins for botulism, smallpox, and anthrax as a result of BioShield.

Presidential directive HSPD-21 (2007) set a new course for public health and medical preparedness in an age of bioterrorism. Instead of a passive strategy whereby doctors and hospitals wait for a medical emergency to arise due to a terrorist or natural event, HSPD-21 advocates a more proactive strategy of bio-surveillance, stockpiling, mass casualty care, and community resilience. The four pillars of this strategy are as follows:

- *National bio-surveillance:* HPSD-21 directs DHHS to develop an epidemiological surveillance system to monitor human disease activity across "state and local government health officials, public and private sector health care institutions, and practicing clinicians … with the principal objective of establishing or enhancing the capabilities of State and local government entities [to provide early warning of disease outbreaks]."
- *Stockpiling and distribution:* Within 48 h of a catastrophic health event, distribute vaccines, drugs, and therapeutics to large populations.
- *Mass casualty care:* Be able to accommodate surge capacity requirements in the event of a healthcare event of "biblical proportions." The directive suggests the use of federal facilities to expand on public and private sector facilities such as hospitals and clinics.
- *Community resilience:* Rely more heavily on local civic leaders, families, and local public health and medical systems—social networks to fall back on—because the federal government cannot do it all.

HSPD-21 may be difficult to implement to its fullest, but it sets a new standard for public health. It is a departure from the traditional Roemer model of public health, which depends heavily on the private sector for medical services, delivery, and research (see Fig. 14.1). Milton I. Roemer (1916–2001), a UCLA public health professor and researcher for 38 years, developed a public health model in 1984 that

[2]Healthcare and Public Health Sector-Specific Plan: An Annex to the National Infrastructure Protection Plan 2010, Department of Homeland Security. www.dhs.gov.

[3]https://en.wikipedia.org/wiki/Project_Bioshield_Act

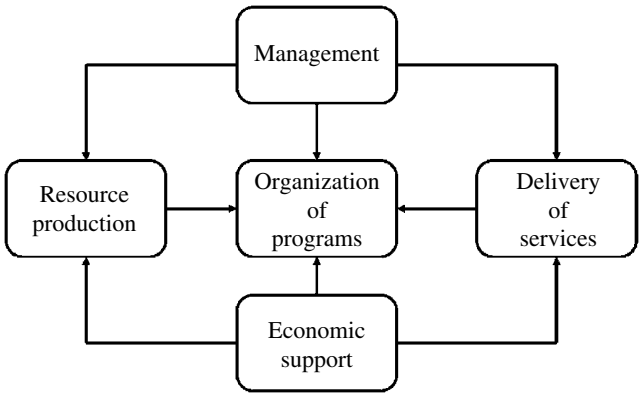


FIGURE 14.1 Roemer's model is a simplified model of the public health sector as it is practiced in the United States.

survives today as a unified big picture of the complex public health sector. Bio-surveillance, stockpiling, mass casualty surge capacity, and community resilience are conspicuously absent in Roemer's model.

14.2.1 Components of Roemer's Model

Figure 14.1 contains five major components of the HPH sector: management, organization of programs, resource production, economic support, and delivery of services. The US implementation of Roemer's model is vertically distributed across federal, state, local, and tribal jurisdictions and horizontally distributed across public and private sectors. However, the private sector is by far the largest piece in the puzzle. This structure poses challenges to the management component, consisting of:

- Planning
- Administration
- Legislation
- Regulation

The organization component consists of the following departments, agencies, corporations, and government funding programs:

- DHHS, including the CDC, Food and Drug Administration (FDA), National Institutes of Health (NIH).
- Private sector entities such as medical insurance companies, pharmaceutical companies, and drug stores.
- Medicare/Medicaid, which supplies nearly 1/3 of funding.
- State, local, and tribal departments of public health, which collect and store vital statistics (birth/death certificates), provide environmental and sanitation oversight, provide testing labs, and monitor and prevent diseases at the local level.
- Voluntary organizations such as the Red Cross, March of Dimes, AARP, and American Medical Association (AMA).

This component is further complicated by dense interdependencies among partners: The Department of Defense (DoD), Department of Veteran Affairs, Department of Justice, Drug Enforcement Agency, Department of Agriculture, Department of Labor, and Environmental Protection Agency (EPA).

The resource production component is even more complex in the US system, because of its interactions with all other components. It consists of:

- Healthcare professionals including 900,000 physicians and 2 million nurses, pharmacists, and so on.
- Hospitals, pharmaceutical companies, and related research organizations. In the United States, hospitals consume 33% of financial resources: physicians 23; drug companies 10%, and biomedical research organizations 3%.
- The pharmaceutical industry that does research and drug development and delivery.

The economic support component provides basic funding for the HPH sector, including:

- Personal households that pay 15% of medical expenses out of pocket.
- Private health insurance companies that pay 34% from policies.
- Social security and other government support: 17% from Medicare and 16% from Medicaid in 2005.

The impact of the Affordable Care Act (2010) is not known at this time, because it did not begin to roll out until 2014. However, the ACA will likely have a major impact on shaping the HPH sector over the long term.

The delivery of services component consists of primary care, long-term care, hospice, and mental health treatment organizations and programs. At the time of this writing, US Medicare pays for 80% of delivery of services. It plays a major role in the healthcare system, accounting for 21% of total national healthcare spending in 2012, 28% of spending on hospital care, and 24% of spending on physician services.

Medicare benefit payments totaled $536 billion in 2012: roughly two-thirds for part A (hospital) and part B (physician), 20% for part C, and 10% for part D (drugs). Medicare is funded from three primary sources: general revenues (40%), payroll tax contributions (38%), and premiums (13%). It is projected to cost $1.1 trillion by 2023, because of an aging population and rising prices.

Funding is the primary threat facing this sector. Government spending on healthcare is projected to double over the next 15 years—from roughly 7% of GDP to 14%. The consequences of this rapid expansion are unknown. If it does not crowd out other programs, and governmental support declines, consumers will have to accept fewer services and healthcare benefits. The risks of public health sector failure come from within.

14.3 THE COMPLEXITY OF PUBLIC HEALTH

The US HPH sector is complex, which means it is subject to complex systems dynamics. For example, is it at risk due to the tragedy of the commons? Might it be another example of the paradox of enrichment? The public health infrastructure is an industrial commons with predators in the form of an aging population, insurance companies, and professional healthcare workers. These predators depend on a CIKR under financial and political pressure unlike any other sector. What happens when nonlinearities like a disaster or bad legislation is introduced into this complex CIKR? Does it collapse or does it adapt?

The following three scenarios are hypothetical, of course, but they are motivated by the historical and projected data shown in Figure 14.2. Government spending has steadily risen from 16% of GDP in 1980 to over 45% in 2013. Similarly, healthcare spending as a percentage of GDP has risen from 1.5% to nearly 9% in 2013. Projections to 2030 are based on these growth curves and predict an eventual financial crossroads in the future. Either technical or political innovations will be needed to bend these cost curves, or government and healthcare services will eventually need to be curtailed. Healthcare costs cannot exceed 100% of GDP or government spending.

Scenario #1: Tragedy of Commons. In this scenario predators are beneficiaries of government-sponsored healthcare such as Medicare and Medicaid at the federal level and city and county public health departments at the local level. Recall that the commons becomes unsustainable if predators act in their own selfish interest and ignore excesses that may collapse the commons. When the cost of medical procedures is disconnected from premiums, consumers will use all of the medical services they can without regard to cost. By optimizing their benefits, consumers are healthier and live longer lives. But increased use and longer lives exacerbate the load placed on the commons. Eventually, the *carrying capacity* of the healthcare commons is exceeded, and the CIKR collapses. Under this scenario, the HPH sector becomes unsustainable sometime during the next 20–30 years. Figure 14.2 shows the steadily increasing cost of government and healthcare since 1980. Can the monotonic increase continue forever?

Scenario #2: Paradox of Enrichment. This scenario is related to the first scenario with the addition of an enrichment clause. Drug manufacturers, hospitals, and insurance companies benefit from spiraling costs—perhaps fueled by generous payments from Medicare, Medicaid, and personal households. If the rewards are rich enough, these predators expand their businesses until they exceed the carrying capacity of the commons. Rapid rises in government spending attract expansion of the private sector, which in turn fuels more funding, in a spiraling bubble that eventually bursts. Once again, the commons becomes unstable due to excess profits and expanded services that cannot be sustained when the inevitable economic slowdown

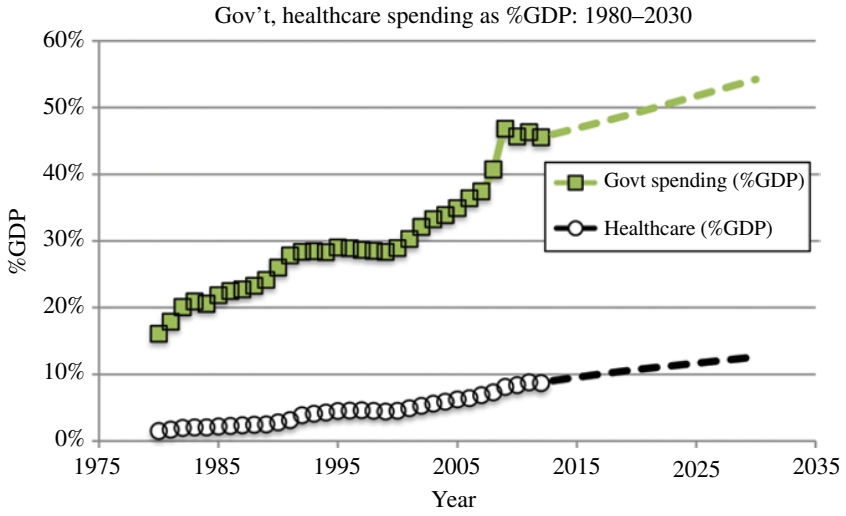


FIGURE 14.2 State, local, and federal government and healthcare spending is expected to steadily rise as a percentage of gross domestic product (GDP).

occurs. Therefore, the HPH sector either contracts or collapses due to an embarrassment of private sector riches—akin to the collapse in housing that occurred in 2008–2009.

Scenario #3: Competitive Exclusion. Insurance companies, like many other infrastructure companies, benefit from *increasing returns*—the more subscribers they have, the more profitable they are. Unit costs decline as basic infrastructure costs are amortized over more and more premium payers. Consolidation of the industrial commons under a near-monopoly is a quick way to achieve massive returns and massive profits. Additionally, cost savings can be applied toward reducing costs to government—a motivation for policy makers to advocate monopoly power. Politicians are likely to turn a blind eye to monopolies that promise to reduce government spending. A similar argument can be made for hospitals, clinics, and pharmaceutical companies. Thus, Gauss's law takes over, and the country ends up with a highly concentrated private sector in control of healthcare. Does monopolistic concentration lead to "too big to fail" criticality?

The foregoing scenarios are speculation, of course, but they illustrate system thinking as a guide for policy makers to consider. What policies might planners and leaders employ to avoid these complexity factors? Moreover, what will political leaders do when the projections in Figure 14.2 are realized?

14.4 RISK ANALYSIS OF HPH SECTOR

The HPH sector is so large and complex that it is difficult to apply risk analysis to the entire sector. It faces threats from terrorism, lack of funding, political disruption, and crushing demands on its people and equipment. Readiness is its key objective, and so it is appropriate to measure its capability against readiness metrics. HPH readiness is measured by its ability to respond quickly (the mobility factor), its sustainability as mentioned above (self-sustainability), and its ability to maintain equipment and people over long periods of time in between catastrophic events (operational readiness).

Figure 14.3 condenses these objectives into a fault tree proposed by Mayer [1]. Mayer defined readiness of this sector in terms of three measurable components: mobility, self-sufficiency, and operations. Each of these components is subject to four vulnerabilities: personnel readiness or lack of it, training, supplies, and equipment. Mayer assumed threats were always present (100%), and vulnerability of each component in Figure 14.3 is the complement of readiness. Therefore, if personnel readiness is 64%, then personnel vulnerability is 76%. Table 14.1 shows the values calculated by Mayer to populate the fault tree in Figure 14.3.

Mayer associated a cost with each readiness factor that fell below 100%. These costs are shown in Table 14.1 as elimination costs. Consequence is the same for all vulnerabilities, because failure in any component means overall failure of the sector. In fact, this is why the risk and vulnerability curves in Figure 14.3 are initially set at 100%.

Mayer found that relatively modest investments in readiness paid dividends in terms of risk reduction, but vulnerability remained high due to the OR-gate logic in the fault tree. The likelihood of system failure is high, but the cost of reducing risk is relatively low (Fig. 14.4). An investment of $100 million, as shown in Table 14.1, reduces vulnerability to less than 50%. The largest investment should be allocated to mobility equipment and personnel readiness, with self-sufficiency personnel coming in third. This result underscores, once again, the criticality of people and funding to the HPH sector.

14.5 BIOTERRORISM

Homeland security in general and critical infrastructure protection in particular is concerned with *bioterrorism*—defined as the unlawful use, or threatened use, of microorganisms or toxins to produce death or disease in humans, animals, or plants. Its purpose is to create fear and intimidation for political, religious, or ideological objectives. Of course, it is not a new threat. Biological warfare has a long and horrible history going back to the 1300s (Siege of Kaffa), the French and Indian War in 1763, and World War I. The use of chemical and biological weapons was banned in 1925 by European nations after witnessing the devastation of World War I. President Nixon ended US research into biological and chemical weapons in 1969. And yet, biological warfare raised its ugly head again during the Syrian revolution in 2012.

Regardless of international agreement by 164 nations to halt research into biological warfare, biological weapon research continued. In 1989–1992 Soviet Union scientists involved in biological weapons research defected to the United States, and we learned that the USSR accidentally released weaponized anthrax spores from its bio-weapons research center in 1979. Chemical weapons were used on Iraqi citizens in the 1980s by dictator Saddam Hussein (Kurdish Genocide in 1986 and Halabja chemical attack in 1988). The *subway sarin attack* in Tokyo on March 20, 1995, killed 8 people and seriously injured 275 others.

Perhaps the largest bioterrorism act in US history occurred in 1984 when members of the Rajneesh cult contaminated an Oregon salad bar with *Salmonella typhimurium* in an effort to influence an election. A ricin attack by the Minnesota militia was foiled in 2001, and the US mail system was used to release anthrax in Florida; Washington, DC; New York; and New Jersey—killing five people.

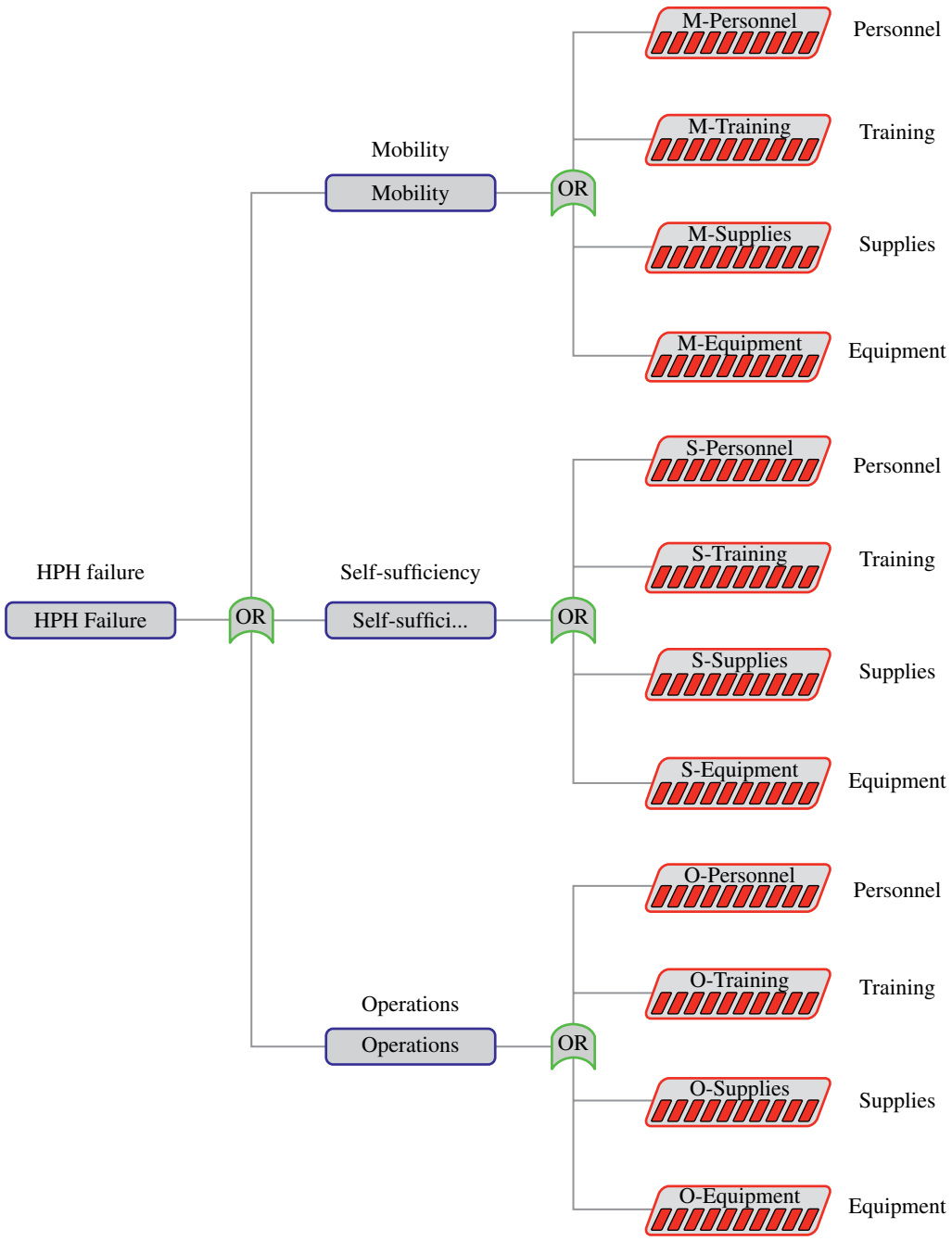


FIGURE 14.3 Fault tree analysis of public health sector risk focuses on threats to mobility, self-sufficiency, and operations: personnel readiness, training, supplies, and equipment.

14.5.1 Classification of Biological Agents

Biological *pathogens* (disease-causing agents) are classified according to their size and biological mechanism:

- *Bacteria:* Single-cell organisms such as anthrax, tularemia, and plague.
- *Rickettsiae:* Parasites that reproduce inside cells.
- *Virus:* Parasites that can only live inside of cells.
- *Fungus:* Pathogens like mold, yeast, and mushrooms.
- *Toxins:* Poison extracted from snakes, insects, spiders, marine organisms, plants, bacteria, fungi, and animals.

The CDC classifies these agents according to their potential impact on public health. *Category A* agents are the most threatening because they can be easily transmitted and disseminated and may result in high fatalities. Examples are tularemia,

TABLE 14.1 An investment of $100 million reduces risk to less than 10% and focuses on mobility equipment, mobility personnel readiness, and self-sufficiency personnel improvements, according to MBRA optimization

Name	Threat (%)	Vulnerability (%)	Elimination cost $(millions)	Consequence $(millions)	Risk initial	Allocation $(millions)	Vulnerability reduced (%)	Risk reduced
M-equipment	100.00	81.00	150.20	1000.00	810.00	43.73	22.53	225.32
M-personnel	100.00	26.00	70.50	1000.00	260.00	12.27	14.75	147.49
S-personnel	100.00	12.00	18.70	1000.00	120.00	6.34	5.17	51.70
O-personnel	100.00	12.00	18.70	1000.00	120.00	6.58	5.00	50.03
M-training	100.00	26.00	21.50	1000.00	260.00	11.86	4.31	43.12
M-supplies	100.00	88.00	11.50	1000.00	880.00	10.23	1.64	16.37
S-training	100.00	24.00	1.20	1000.00	240.00	1.20	1.00	10.00
S-equipment	100.00	78.00	0.17	1000.00	780.00	0.17	1.00	10.00
O-training	100.00	24.00	1.20	1000.00	240.00	1.20	1.00	10.00
S-supplies	100.00	87.00	3.10	1000.00	870.00	3.10	1.00	10.00
O-supplies	100.00	87.00	3.10	1000.00	870.00	3.10	1.00	10.00
O-equipment	100.00	88.00	0.22	1000.00	880.00	0.22	1.00	10.00

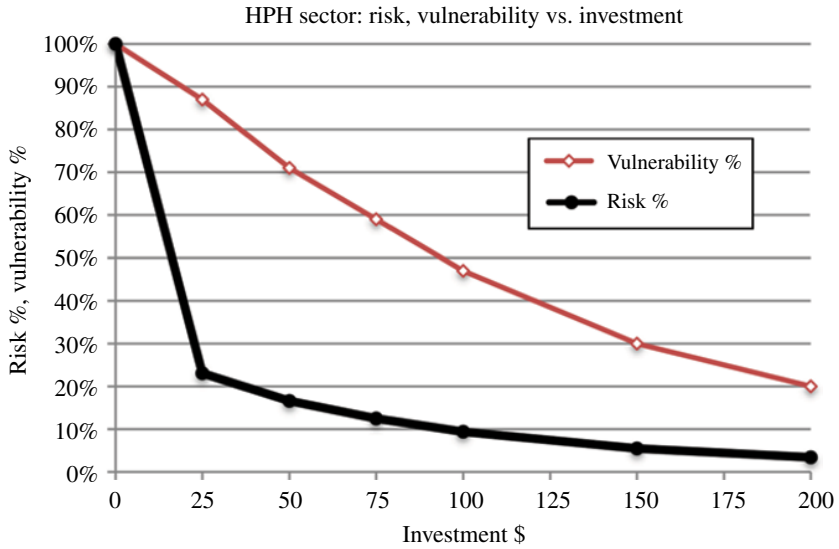


FIGURE 14.4 A relatively small investment dramatically reduces risk because vulnerability reduction is relatively inexpensive.

anthrax, smallpox, botulinum, bubonic plague, and hemorrhagic fevers. They are typically contracted through contact, inhalation, or drinking water. Some are more infectious than others, and most have no cure.

Category B agents are moderately easy to disseminate and have low mortality rates. Examples are brucellosis, *Clostridium perfringens*, salmonella, *E. coli*, glanders, Q fever, ricin from castor beans, typhus, viral encephalitis, and *Cryptosporidium*. The city of Milwaukee was contaminated with cryptosporidium in 1993 when the water treatment plant failed to filter out bacterium from drinking water. Within two weeks, 403,000 residents in the Milwaukee area became ill, and 104 people died.

Category C agents are pathogens that might be engineered for mass dissemination because of their availability, ease of production and dissemination, high mortality rate, or ability to cause a major health impact. Examples are the hanta virus, SARS, the H1N1 or H5N1 strains of influenza, and HIV/AIDS.

In addition to these natural pathogens, a new breed of biological threats is emerging based on *synthetic biology*—the manufacture of organic parts, devices, and systems by genetic engineering. *Synbio*, as it is called, is a new field of bioengineering that combines molecular biology, various fields of engineering, and bioinformatics. White-hats claim that its goal is to create energy, produce food, optimize industrial processing, and detect, prevent, and cure diseases. But it also has the potential to become a weapon of black-hats.

According to Tucker and Zilinskas, "The main difference between genetic engineering and synthetic biology is that whereas the former involves the transfer of individual genes

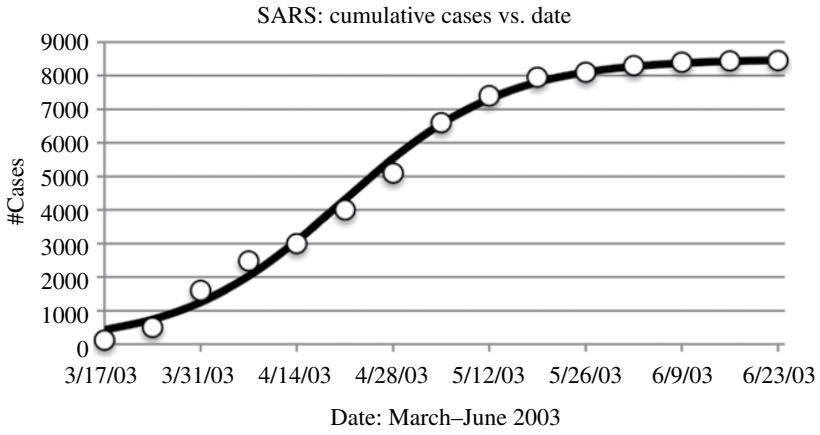


FIGURE 14.5 The number of infected people due to the SARS pandemic obeyed a logistics S-curve predicted by the Kermack–McKendrick model.

from one species to another, the latter envisions the assembly of novel microbial genomes from a set of standardized genetic parts. These components may be natural genes that are being applied for a new purpose, natural genes that have been redesigned to function more efficiently, or artificial genes that have been designed and synthesized from scratch" [2].

For example, Washington University researchers synthesized the hepatitis C virus genome from chemically synthesized parts in 2000, and SUNY Stony Brook researchers synthesized the base poliovirus genome from its published sequence, producing the second synthetic genome in 2002. Synthetic DNA that behaves like natural DNA, made of nonorganic base pairs, has been demonstrated in the lab.[4] How difficult is it to construct an artificial virus that spreads like a highly contagious disease and has no cure?

14.6 EPIDEMIOLOGY

Homeland security is predominantly concerned with catastrophic events such as terrorist attacks and pandemics. An epidemic is an outbreak of a disease among a certain population. A pandemic is an epidemic that spreads over large regions—typically across borders. Pandemics like SARS and H1N1 infect thousands of people in dozens of countries and threaten to spread like a wildfire throughout the entire human population. It is important to understand the dynamics of pandemics because consequences can be enormous even though the probability of occurrence is extremely low. This type of black swan event is so rare that its probability approaches zero as its consequence approaches infinity. Does this place pandemics in the high-risk category?

[4]http://syntheticbiology.org/

14.6.1 The Kermack–McKendrick Model

The scientific study of epidemics is a robust and sophisticated subject that began in 1927 with the first mathematical model devised by Kermack and McKendrick [3]. The *Kermack–McKendrick model* is expressed in four mathematical equations, but they are summarized here in plain English. The four equations represent the state of human individuals in a targeted population under the threat of an epidemic. The target population is divided into groups representing each of the four states—(1) *susceptible* if the individual is available to become infected, (2) *infected* if the individual is ill, (3) *recovered* if the individual has been infected and recovered, and (4) *removed* if the individual has died because of the disease. Individuals move from one state to the other according to the Kermack–McKendrick rate equations [4]:

Susceptible: The rate of change of the number of susceptible individuals declines proportional to infectiousness, number of susceptible individuals, and number of infected individuals. This says the number of susceptible individuals depends on both the number of infected and non-infected individuals.

Infected: The rate of change of the number of infected individuals increases proportional to the difference between the rate of change of susceptible individuals and the number of infected individuals removed. This says the number of infected individuals grows at a rate that depends on both the number of infected and non-infected individuals.

Removed: The rate of change of the number of removed (deceased) individuals is proportional to the number of infected individuals. This says the number of individuals that die from the contamination depends on the number infected.

Conservation: The number of susceptible, infected, and removed must sum to 100% of the target population.

The mathematics underlying these rules produces an S-shaped curve of accumulated number of infected people versus elapsed time since the breakout (see Fig. 14.5). This curve begins with

an *index* person—the first person to be infected—and ends with the last person to contract the disease. It represents the total number of people reported to have contracted the disease. (A similar but different curve gives the number of deaths.)

14.6.2 SARS

SARS is a useful case study in pandemic theory because we know a lot about it. (The following are excerpted from a variety of sources) [5]. The pandemic started in Fushan, China—a province near Hong Kong—in November 2002. It began with dinner ingredients purchased by the first victim on his way home from work—the *index* person. The stir-fried chicken, domestic cat (civet), snake, and vegetable dinner became internationally infamous after the index person and five others died a few days later. An investigation by public health experts from around the globe traced the cause of death to the civet—a catlike animal infected with a previously unknown virus called severe acute respiratory syndrome (SARS). The SARS virus probably originated in bats that then passed it on to civets and then on to humans.

Later, a stir-fry cook living in Heyuan—560 miles from Fushan—checked into a local hospital with flu-like symptoms that developed into an illness so severe that he was transferred to Guangzhou 2 days later. Doctor Liu Jianlun, a physician who never suspected what he was up against, accompanied the cook on his journey. The doctor attended a wedding in Kowloon, 102 miles from Heyuan, and checked into room 911 of the Kowloon Metropole Hotel. He inadvertently infected six or seven people while waiting for the lift on the ninth floor. Next to him were three Canadians, a man from Hong Kong, an American businessman, and a woman from Hong Kong.

The people waiting for the lift formed a social network that began to expand as people traveled around the globe. The first generation of this social network became part of the *base reproduction number* R_0—the average number of people infected by coming into contact with an infected person. For SARS, R_0 is thought to equal about 4–6 people (2.7 is generally used) (see Fig. 14.6 for the complete social network formed by the spread of SARS from China to 29 other countries.)

On February 21, 2003, Doctor Jianlun became ill and checked out of his room and went to a Hong Kong hotel. By March 4, 2003, he was dead. Two of the Canadians waiting for the lift checked out of the hotel and spent the night with their son in Hong Kong. Later they flew home to Canada and infected their immediate family members. In fact, everyone waiting for the lift with Jianlun became ill and hospitalized in various countries, including China, Vietnam, Singapore, and Canada. Additionally three women from Singapore who occupied a room on the ninth floor returned home and became ill. Another guest of the hotel, Johnny Cheng, carried the disease to Hanoi and died after spreading the contagion further.

Doctor Carlo Urbani, an infectious diseases specialist working for the World Health Organization (WHO) in Hanoi, attended Johnny Cheng. Urbani noticed the spread of flu-like symptoms among hospital workers and initially suspected it was avian influenza but later realized it was something new. Unfortunately, he had already contracted the new disease. On March 11 he flew to Bangkok to attend a medical conference and died there on March 29, 2003.

The emerging social network eventually spread SARS around the globe, infecting 8422 and killing 916 people in 29 countries before it dissipated and faded by June 2003. Guests panicked and cleared out of the 487-room Metropole Hotel when its role as SARS hub was announced on March 19, 2003. Even though the contagion began on November 16, 2002, the People's Republic of China delayed notifying the WHO until February 10, 2003—3 months, 806 cases, and 34 deaths later. After considerable international criticism the country's Health Minister apologized for delays in reporting, and Chinese medical officials began reporting the status of the SARS regularly on April 2, 2003. Figure 14.5 was obtained from these reports and others.

Perhaps more interesting than the near-perfect fit to the Kermack–McKendrick epidemic curve is the question, "Why did SARS stop?" That is, why did only 916 people die and the contagion stop after reaching only 29 countries? After all, modern air travel is supposed to spread pandemics around the globe, making pandemics one of the highest high-risk threats to humanity. In theory, SARS could have contaminated millions of people and killed 10% of those contaminated. The Spanish influenza (H1N1) of 1918–1920 infected 500 million (30–50% of the global population), killing 50–100 million. World War I, which ended the same year, claimed 9 million lives.

14.7 PREDICTING PANDEMICS

One of the primary objectives of HPH is to detect pandemics early on and prevent their global spread. The halt of SARS should be considered a major victory for the HPH sector, because within a matter of months, SARS was stopped dead in its tracks. Figure 14.7 shows what happened in the United States. SARS was spreading according to the Kermack–McKendrick S-curve until healthcare professionals responded and resoundingly stopped it in a matter of days. Experts generally agree that SARS could have spread to every corner of the world and killed millions of people. Why was it quickly stopped?

The traditional theory of countermeasures recommends a *blocking strategy* whereby a fraction of the population κ_B is inoculated or quarantined to remove them from harm's way. The fraction of the population that needs to be inoculated or quarantined to block the spread depends on the base reproduction number R_0 and a speedy response:

$$\kappa_B = 1 - 1/R_0$$
$$\therefore \kappa_B(SARS) = 1 - 1/2.7 = 0.63$$

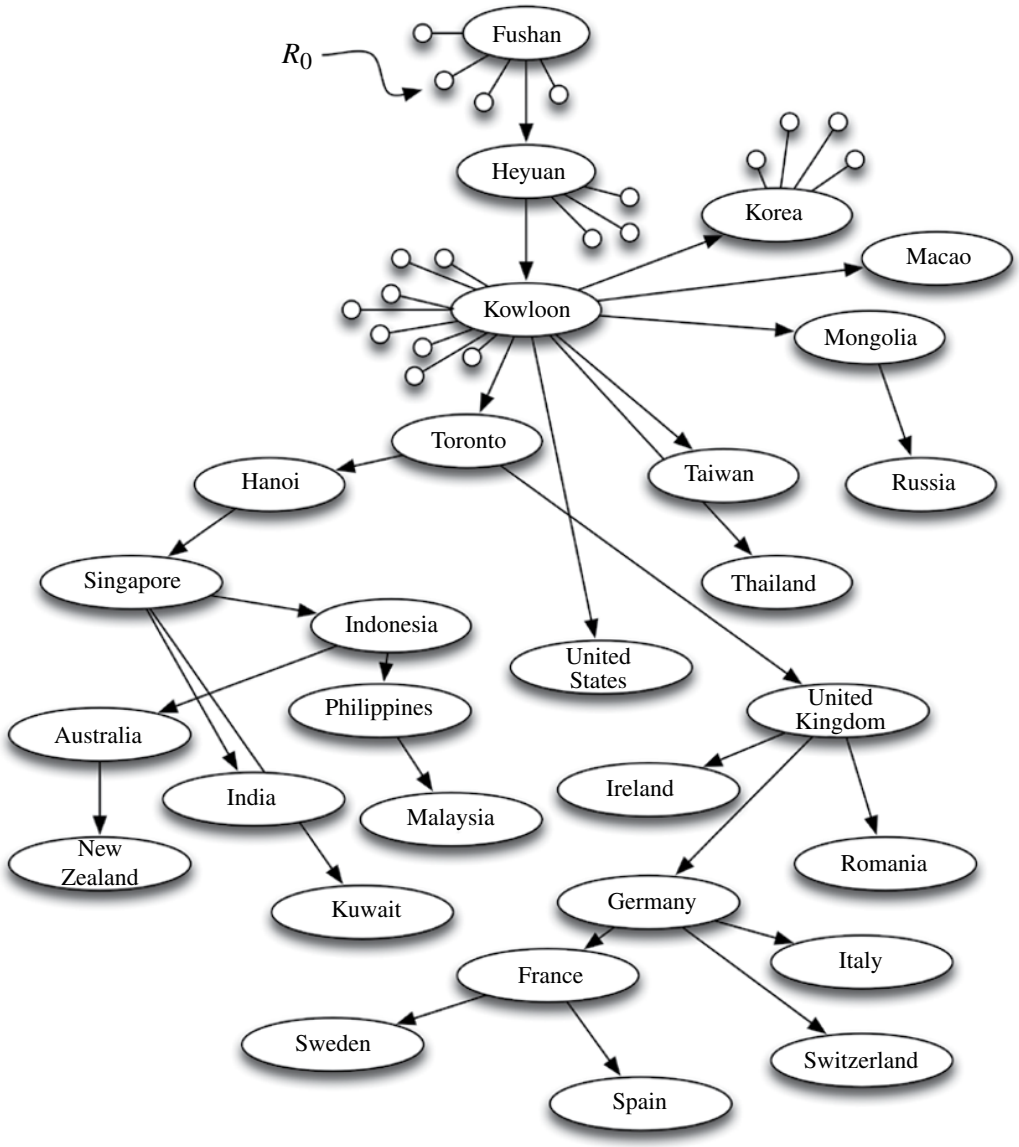


FIGURE 14.6 The SARS pandemic formed a social network that spanned 29 countries. The distance between subsequent outbreaks forms a spatial Levy flight with fractal dimension of approximately 1.6.

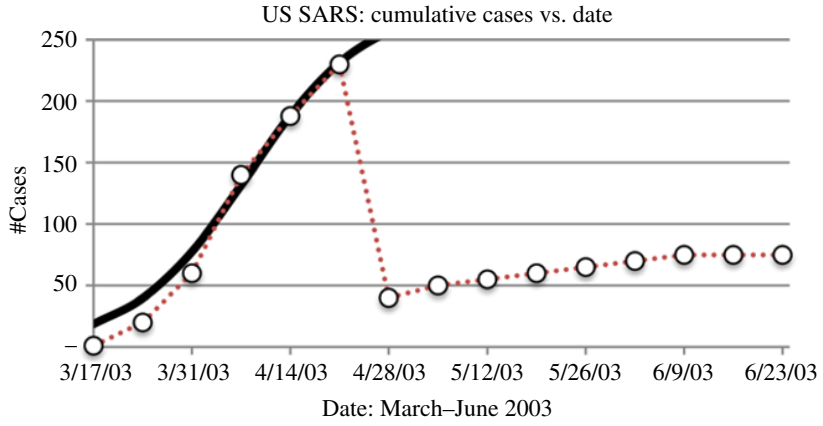


FIGURE 14.7 The spread of SARS within the United States and globally was abruptly halted by quick-acting healthcare professionals.

The global healthcare commons would have had to inoculate or isolate 63% of the world's population to implement this strategy. It is highly unlikely that (0.63)(7 billion) = 4.4 billion people could be inoculated or isolated in time even if funding was available. Of course, it is not necessary to treat everyone if the epidemic is contained within a single region. For example, the nearest 250 million people within a few thousand miles of Hong Kong might have been adequate. But treating (0.63)(250 million) = 157 million people is still a daunting task. The traditional strategy of blocking a large percentage of the population is inadequate when it comes to pandemics.

14.7.1 The Levy Flight Theory of Pandemics

Why did the bubonic plague kill one-third of the European population in the fourteenth century and yet SARS died out after contaminating thousands and killing hundreds? Fourteenth-century Europe lacked globe-hopping air travel, so the modern air transportation system should make pandemics worse. In fact, it appears that commercial air travel spread SARS beyond China's borders. But at least one group of researchers argues that air travel mitigates the spread. SARS was stopped because of air travel and quick-acting public health authorities.

There are two fundamental kinds of epidemics. The first kind, *susceptible–infected–recovered* (SIR), describes individuals in a population that are initially susceptible to a disease, then infected with a certain probability, and finally either removed or recovered so they are no longer susceptible (see Fig. 14.8a). An individual starts out in the susceptible state and transitions to the infected state with probability γ and either dies with probability Δ or recovers with probability $(1 - \Delta)$.

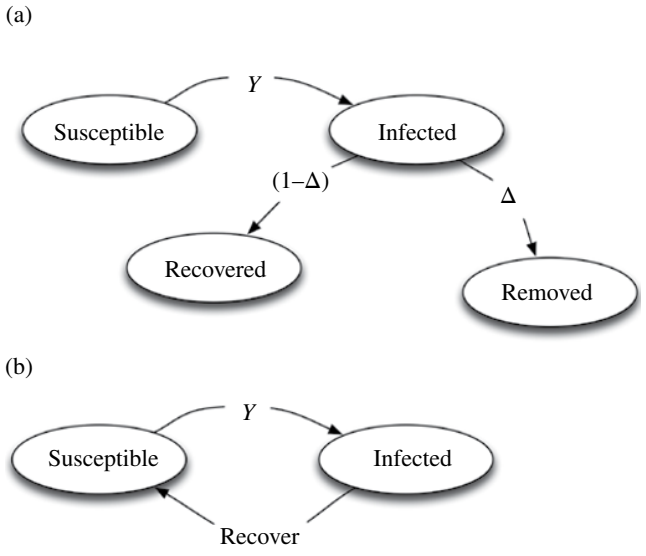


FIGURE 14.8 State diagrams of SIR and SIS epidemics differ—under certain conditions, SIS epidemics can persist. (a) SIR model. (b) SIS model.

SIR describes many biological populations such as humans that become immune to subsequent infections after recovering. SIR diseases eventually die out because they either kill their hosts or become immune to repeated infection. Interestingly, it is often unclear whether a disease is eradicated forever, because it may break out again after lying dormant for years. There have been no new outbreaks of SARS since 2004, and smallpox was eradicated in the 1970s.

The second kind of disease is known as *susceptible–infected-susceptible* (SIS). It spreads throughout a population of individuals that are initially susceptible, then infected with a certain probability, and finally recover, only to become susceptible, again (see Fig. 14.8b). An SIS population can die out or sustain a contagion forever, if conditions are right for recurrence of the disease. If it never completely dies out, it is considered *persistent*. Some Internet viruses are persistent—waxing and waning, but never vanishing completely. Once again, it is sometimes difficult to determine if a disease is persistent or not. Bubonic plague (black death) may be persistent. In 2012, 60 people in Madagascar died of the plague, and contaminations were reported in the United States in 2013.

We know from earlier chapters that a complex CIKR network will support a persistent SIS virus if the product of infectiousness and spectral radius exceeds the rate of recovery:

$$\text{Persistent} : \gamma\rho > \Delta$$
$$\Delta : \text{removal rate}$$

When $\gamma\rho$ is much greater than Δ, cascade failures in CIKR networks transition from high risk to complex catastrophic risk. The exceedence probability divides into two modes—a long-tailed mode as expected, plus a binomial distribution mode representing catastrophic risk. In social networks where individuals recover from an illness, a certain fraction of the recovered population becomes infected again before the disease can be entirely eradicated. The contagion is never completely eliminated. It is SIS and persistent.

Understanding persistence is an important skill in the war against Internet viruses and eradication of virulent diseases in fixed populations. The theory helps to fight the Internet virus or human disease. But the theory falls short of completely explaining diseases like SARS and H1N1 in a global social network that expands across the world. To understand persistence—or the lack of it—in global diseases, we need a spatial theory. As it turns out, *Levy flights* are a perfect match, because Levy flights in spatial networks like the one shown in Figure 14.6 may explain why SARS suddenly vanished after only 6–8 months.

As it turns out, the distances traveled by the SARS virus obeyed a power law with fractal dimension of approximately 1.6. This number is less than 2.0—the *critical point* separating a pandemic that covers the globe from a pandemic that dies

out. That is, the power law obtained from tracing a Levy flight pandemic like SARS determines whether the disease continues to spread until reaching everyone on the planet, or not.

In 2010 a group of researchers studying the spread of SARS and H1N1, led by Yanqing Hu of Beijing Normal University, found a relationship between Levy flights and SIS spreading that may explain why SARS stopped [6]. They argued that an epidemics like SARS and H1N1 die out if the power law fractal dimension is less than or equal to 2, claiming, "the epidemic is liable to disappear if there are more long-distance connections than short ones." SARS may have vanished because it jumped too far each time it spread.

Recall that the length of the long-tailed power law decreases as the fractal dimension increases. They are inversely related. Therefore, as the distances separating outbreaks decrease, the power law dimension increases. Short air travel hops produce short-tailed power laws. A Levy flight with dimension 2.0 has shorter air travel hops than one with a dimension of 1.6. So, Hu et al. argue that short hops lead to pandemics that are capable of spreading throughout the world, while long hopping pandemics quickly die out. This is why the bubonic plague nearly wiped out Europe—it took small steps of approximately 3 miles/day. SARS, on the other hand, jumped thousands of miles per day.

Hu's Levy flight explanation uses a conservation of energy argument to explain why contagions die out if they take large steps. But conservation of energy may not apply to biological contamination. Why should a germ conserve energy? A simpler explanation may be that small fires are easier to put out than large fires. Think of an outbreak of SARS as a bonfire. If left alone, one large bonfire can get out of control and consume a large forest, but if divided into sections that are carried off to distant forests, the large fire is reduced to many smaller ones. Add a parallel response—many healthcare workers working simultaneously on smaller outbreaks to put them out—and Occam's razor applies. SARS may have disappeared in a matter of months because of simple divide and conquer and a global HPH infrastructure that responded quickly and effectively.

14.8 BIO-SURVEILLANCE

The SARS experience underscored the importance of early warning and quick responses to potential pandemics. Consequently, a number of bio-surveillance services have been created to collect early warning signs of impending outbreaks throughout the world. Monthly reports of outbreaks can be seen at www.healthmap.org, for example, and travelers can look up disease activity at GeoSentinel, located online at http://www.istm.org/geosentinel/main.html.

14.8.1 HealthMap

According to its Web site, "HealthMap, [is] a team of researchers, epidemiologists and software developers at Boston Children's Hospital founded in 2006… [and] brings together disparate data sources, including online news aggregators, eyewitness reports, expert-curated discussions and validated official reports, to achieve a unified and comprehensive view of the current global state of infectious diseases and their effect on human and animal health. Through an automated process, updating 24/7/365, the system monitors, organizes, integrates, filters, visualizes and disseminates online information about emerging diseases in nine languages, facilitating early detection of global public health threats."[5]

Outbreaks are shown on a map along with details of the nature of the contagion, how many humans and animals were affected, and any other circumstances. For example, during the week of October 21, 2013, HealthMap reported, "10 Gastrointestinal Alerts: Salmonella (8), E. coli (1), Waterborne Illness (1)" near Salinas, California. The service is limited, however, to weekly, monthly, and annual summaries. There are no time-series data to analyze.

14.8.2 Big Data

A number of nongovernment sponsored Web sites use *big data analytics* to track and report outbreaks. For example, www.SickWeather.com animates real-time data obtained from social network sites like Facebook.com and Twitter.com. After analyzing 17 million mentions of illness in Facebook.com posts and Twitter.com tweets, Sickweather founder Graham Dodge noticed that disease spreads most quickly between Hartford, Connecticut, and Washington, DC, a corridor he called "contagion alley."[6]

SickWeather.com scans social networks for mentions of 24 different symptoms. It then automatically separates casual posts from posts that actually refer to illness and plots them on a map, using semantic analysis (algorithms for extracting meaning from sentences). Consumers can look up a disease on a map and track its geographic spread over time.

Malaria kills over 650,000 people every year, so eradication of malaria is one of many aims of global healthcare agencies, including the UN. A big data study led by researchers at Harvard School of Public Health and seven other institutions correlated the movement of malaria-infected people with GPS traces from cell phone calls [7]. They found that malaria emanates from the mosquito-infected Lake Victoria region and spreads east to Nairobi, where it is accelerated by the high concentration of people in

[5]http://healthmap.org/site/about
[6]http://mashable.com/2012/06/08/social-media-disease-tracking/

Nairobi. After contracting the disease, the infected cell phone-toting shoppers and merchants return to their villages where they infect others.

The researchers mapped every call or text made by 14,816,521 Kenyan mobile phone subscribers to one of 11,920 cell towers in 692 different settlements. They were able to identify the flow of malaria through the big data collected by cell phones. The result of cellular phone bio-surveillance was turned into actionable policies to fight malaria.

14.8.3 GeoSentinel

GeoSentinel is a worldwide communication and data collection network for the surveillance of travel-related diseases. The *International Society of Travel Medicine* (ISTM) and the *Centers for Disease Control* (CDC) created GeoSentinel in 1995, to aggregate data from 57 globally dispersed medical clinics. The Morbidity and Mortality Weekly Reports (MMWR) contain impressive information. For example, the July 19, 2013, MMWR contains summary statistics for the period 1997–2011:

- Commercial aviation and international civilian travel has increased steadily to 1 billion trips in 2012. Tourism comprises approximately 5% of the total worldwide gross domestic product, with growth coming largely from emerging economies.
- In 2009, US residents made 61 million overnight trips outside the country, including 14% of US students pursuing a bachelor's degree abroad. In 2011, one-third of 27 million US residents traveling overseas listed visiting friends and/or family as their main reason to travel. This includes immigrants and their children who return to their country of origin to visit friends and relatives (VFR travelers).
- US residents spent $79.1 billion on international tourism in 2011. This represented 7.7% of the world's international tourism market, making the United States second only to Germany in terms of the international tourism market share.
- From September 1997 to December 2011, 164,378 patients were surveyed and included in the GeoSentinel Surveillance System's database. 141,789 (86%) reported probable travel-related illness. Included in this analysis were 10,032 after-travel patients with 13,059 confirmed or probable final diagnoses (1.3 diagnoses/patient).
- The most frequent diagnosis was *Plasmodium falciparum* malaria. Approximately half of patients who contracted *P. falciparum* malaria in Sub-Saharan Africa were visiting friends or relatives. Proper malaria chemoprophylaxis and mosquito bite avoidance should remain a priority.

In addition to concern over malaria, GeoSentinel lists 60 other agents responsible for illnesses due to travel. It is the largest repository of provider-based data on travel-related illness. But, for the United States, the risk of death due to travel-related diseases and accidents is far less than deaths due to everyday mundane illnesses. Table 14.2 lists the top causes of death in the United States and the top factors that contribute to deaths. Clearly, smoking and diet/exercise is a bigger threat than bioterrorism or global pandemics. In fact, researchers at University of South

TABLE 14.2 The major causes of death as reported on a death certificate are not the same as the major reasons why people die of heart disease and cancer[a]

Cause of death (2000)	#Cause	%Cause	Factor	#Factor	%Factor
Heart disease	615,651	25.4	Tobacco	435,000	17.9
Cancer	560,187	23.1	Diet/exercise	400,000	16.5
Stroke (brain)	133,990	5.5	Alcohol	85,000	3.5
Lung disease (COPD)	129,311	5.3	Microbial agents	75,000	3.1
Accidents	117,075	4.8	Toxic agents	55,000	2.3
Alzheimer's	74,944	3.1	Car crashes	43,000	1.8
Diabetes	70,905	2.9	Firearms	29,000	1.2
Influenza/pneumonia	52,847	2.2	Sexual behaviors	20,000	0.8
Kidney diseases	46,095	1.9	Illegal drugs	17,000	0.7
Blood poisoning	34,851	1.4	Misc.	1,265,059	52.2
Suicide	33,185	1.4			
Liver disease	28,504	1.2			
Hypertension	23,769	1.0			
Parkinson's	20,136	0.8			
Homicide	17,520	0.7			
Misc.	465,089	19.2			

[a]References [8, 9].

Carolina modeled the likelihood of death by heart disease using a Bayesian belief network and reported 93% likelihood when smoking, poor diet, and lack of exercise are combined [10].

14.9 NETWORK PANDEMICS

The Kermack–McKendrick model makes a major assumption that may not hold in the modern connected world, because it assumes *uniform mixing*—everyone in the target population comes into contact with everyone else with equal probability. People are not like molecules in a room—bouncing off of one another with equal likelihood. Rather, modern societies form social networks containing hubs and betweeners. Some people are comparatively isolated, while others are *super-spreaders*—hubs with many contacts (links) with others. Therefore, a social network model of epidemics is needed in place of the Kermack–McKendrick uniform mixing model.

The *paradox of redundancy* says robustness is increased by percolation, but cascade resilience is decreased. A network with many links may improve its performance, but it also decreases its resilience against cascades. Paradoxically, robustness is a disadvantage when it comes to epidemics. Instead, cascade resilience is improved by de-percolation (removing links) and blocking (removing blocking nodes).

Inoculation and quarantine are the means of de-percolation and blocking, so these are the tools used to combat the spread of an infectious disease. De-percolation of redundant links and blocking by hardening the critically important blocking nodes reduce the spread of a contagion by reducing spectral radius (de-percolation) and blocking (partitioning the social network into islands).

Recall from earlier chapters that link robustness was obtained by counting the number of links that are nonessential to the connectivity of the network. The algorithm is simple but time consuming. Examine each link one at a time. If its removal separates the network into disjoint components, keep the link. Otherwise, remove it. A good approximation of the number removed is $(m - n)$, where n is the number of nodes and m is the number of links in the original network.

Node robustness is obtained in a similar manner. Examine each node one at a time, and if its removal separates the network into disjoint components, keep the node and mark it as a blocking node. A rough approximation to the number of blocking nodes is n/ρ, where ρ is the spectral radius of the network. However, an exhaustive enumeration algorithm is needed to identify which nodes are blocking nodes.

These two techniques replace the Kermack–McKendrick pandemic countermeasure based on the basic reproduction number R_0 and $1 - 1/R_0$. Modern network countermeasures use mean degree λ and *spectral radius* ρ in place of R_0 to reduce the spread of a pandemic. The equations developed in Chapter 4 approximate what is needed—the number of links (contacts) and nodes (individuals) that must be removed by inoculation or quarantine:

$$
\begin{aligned}
&n : \#\,\text{nodes}\\
&m : \#\,\text{links}\\
&\lambda : \text{mean degree}\\
&\kappa_L = 1 - 2/\lambda\\
&\#\,\text{removed links} = m - n;\\
&\kappa_B = 1/\rho\\
&\#\,\text{blocking nodes} = n/\rho
\end{aligned}
$$

Actually, n/ρ equals the number of *first-order blocking nodes* required to partition a social network into disjoint components. An air gap is introduced by removing the blocking nodes, but this serves only to divide the network up into connected neighborhoods. If a disease strikes one individual within a neighborhood, it can still spread to others in the same neighborhood. Blocking reduces cascading, but it does not eliminate it.

As an example, consider the 9/11 terrorist network that attacked the United States on 9/11 shown in Figure 14.9. This network illustrates the impact of removing the following first-order blocking nodes from the terrorist network:

Atta
al Hazmi
Khemais
Beghal
Darkazanli

Assuming node vulnerability of 20%, the fractal dimension of the 9/11 terrorist network is 0.66. It is long tailed and therefore high risk. If the five first-order blocking nodes are removed from spreading the contagion, the fractal dimension increases to 0.95. This reduces risk of $C = 50\%$ or more of the network becoming infected by a factor of three. The calculation is shown here for convenience:

$$\frac{R(0.66)}{R(0.95)} = \frac{C^{1-0.66}}{C^{1-0.95}} = C^{0.29} = 50^{0.29} = 3.1$$

The second-order blocking nodes are:

al Shibh
al Mihdhar
al Alghamdi
Maaroufi
Bensakhria

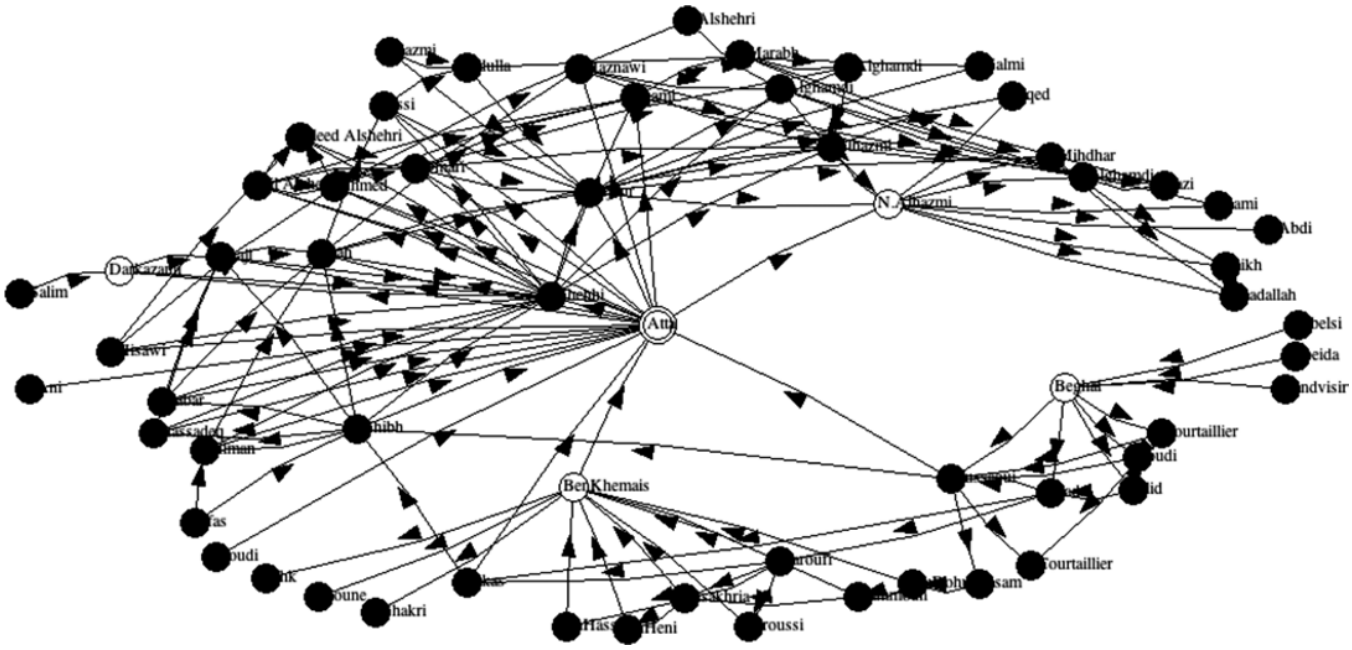


FIGURE 14.9 The four first-order blocking nodes of the 9/11 terrorist social network separate the network into disjoint components. Blocking nodes are white.

Removing more nodes reduces risk further. Each time a blocking node is removed from further spreading, it also removes links. Therefore, blocking de-percolates the network at the same time that it partitions it into disjoint components. This leads to an obvious strategy: inoculate or isolate the first-order blocking nodes, first, followed by second-order nodes, and so on, until the epidemic or pandemic is stopped. The 9/11 terrorist network of 62 nodes contains 16 first-, second-, and third-order blocking nodes. Removing all of them reduces the number of links from 150 to 51 and the number of non-blocking nodes from 62 to 46.

Removing the first-order blocking nodes reduces risk by a factor of three. Removing all blocking nodes reduces risk by a factor of 5. But removal of all blocking nodes has diminishing rewards, and removal becomes expensive. De-percolation may be a more effective strategy, but it is more difficult to implement. Link de-percolation to reduce self-organization removes redundant links. Redundant links are connections that can be removed without separating the network into disjoin components. Removal of all redundant links leaves the network connected so that information (or contagion) can flow from any node to any other node. But de-percolation minimizes cascading—the objective of epidemic countermeasures.

If de-percolation of 59% (89) of the 150 links is combined with blocking, spreading is minimized. Assuming node vulnerability $\gamma = 10\%$, the 9/11 terrorist network yields a fractal dimension of 1.23 without any blocking or de-percolation. Blocking of 5 critical nodes increases fractal dimension to 1.50, and de-percolation of 89 links increases resilience to 2.30. Blocking and de-percolation combined halts the spread entirely. But hardening of 89 links and 16 nodes is expensive.

14.10 THE WORLD TRAVEL NETWORK

The world travel network is the network formed by commercial airline routes (links) and airports (nodes). This network forms a social network of travelers that looks much like the 9/11 terrorist network (see Fig. 14.10). Unfortunately, it is a disease vector that can transmit diseases such as SARS to the far reaches of the globe. Obviously, public health officials are concerned with methods of blocking the spread of contagious diseases through this network. But it is extremely difficult and costly to protect every airport, every flight, and every passenger directly. Is it possible to block the spread of pandemic contagions by protecting a reasonably small subset of the commercial airline's routes and airports? In this section, the OpenFlight network of airports and routes is studied to illustrate the value of blocking and de-percolating to prevent the spread of pandemics.[7]

The spectral radius of the top 500 airports and 4096 routes in the OpenFlight500 network is 45.3, and the spectral radius of the OpenFlight1000 network is 55.8. These are extremely self-organized as suggested by the miniature version shown

[7]http://openflights.org/data.html

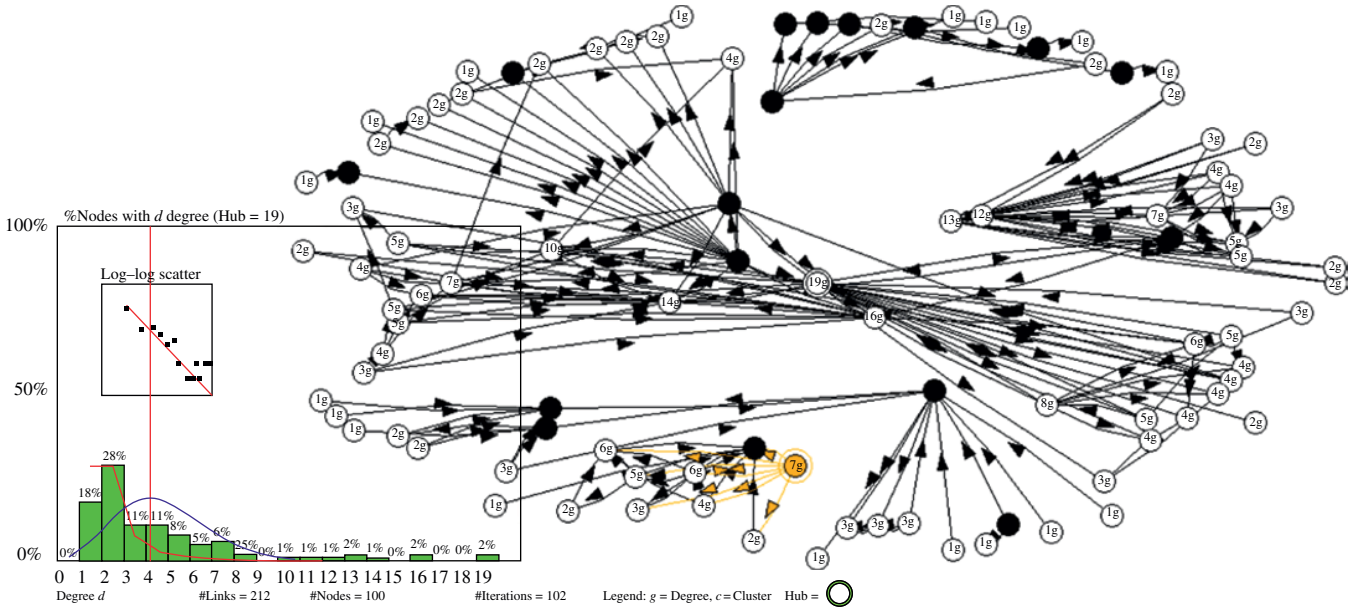


FIGURE 14.10 The OpenFlight100 network shown here with $n = 100$ airports and $m = 212$ routes is scale-free with spectral radius of $\rho = 8.93$. Black nodes are the first-order blocking nodes.

in Figure 14.10. A virus with very low infectiousness should theoretically spread throughout the world via this network. In the following, a vulnerability of $\gamma = 2\%$ is used to obtain fractal dimensions of cascading viruses spreading through this network.

Simulation of the spread of a contagion through an unprotected network yields fractal dimensions of 2.95 and 3.05, respectively, for OpenFlight500 and OpenFlight1000. There are 35 blocking nodes in OpenFlight500 and 84 in OpenFlight1000. By hardening these critical airports, fractal dimension increases to 3.40 and 4.61, respectively. De-percolation of routes eliminates the spreading altogether (assuming $\gamma = 2\%$). But de-percolation of links means cancelation of 88 and 86% routes, respectively. Flights are the revenue generators for the airline industry, so it is unlikely that 88 or 86% of flights will be canceled. Additionally, 40–43% of the airports would have to be blocked to completely eliminate spreading.

The most promising blocking nodes and links in a scale-free network like OpenFlight are the high degree and betweenness nodes. These are super-spreading nodes, because of their connectivity. Therefore, a more practical strategy for reducing the spread of a virulent contagion is to harden the highest degree and betweenness nodes (airports). That is, a select number of airports are designated as super-spreaders. BioShield equipment should be placed throughout these airports to detect contagious diseases prior to boarding and travel.

For OpenFlight500, hardening of the top 10 airports, ranked by normalized degree and betweenness, yields a fractal dimension of 5.75. The top ten airports in OpenFlight500 are:

AMS
FRA
PEK
MNL
IST
ZRH
DME
HKG
JFK
MUC

Hardening the top 20 nearly eliminates all spreading at the 2% level. Blocking these super-spreaders reduces the number of active airports to 480 and the number of routes from 4096 to 2646—a reduction of 35%. This is far less than the blocking strategy that requires blocking of 40% of the airports and 88% of the routes.

The OpenFlight500 simulation suggests a ranking of strategies for large scale-free networks:

- Harden the nodes with the highest value of normalized degree and betweenness first.
- If resilience is not raised to the desired level, identify the blocking nodes next and harden them.
- If more resilience is desired, de-percolate links until an acceptable level of resilience is reached.

14.11 EXERCISES

1. The US healthcare and public health sector is the largest industrial commons in the United States, consuming what percent of US GDP?
   a. 100%
   b. 17%
   c. 45%
   d. 4.5%
   e. 10%
2. Key public health legislation began with:
   a. Food, Drug, and Cosmetic Act of 1938
   b. Affordable Care Act of 2010
   c. Bioterrorism Act of 2002
   d. Homeland Security Act of 2003
   e. Roemer's model of 1984
3. In the United States, public health is predominantly run by:
   a. Federal government agencies
   b. State, local, and tribal agencies
   c. Medicare and Medicaid
   d. Department of Health and Human Services
   e. The private sector
4. HSPD-21 redirected public health, placing more emphasis on:
   a. Community resilience
   b. Bio-surveillance
   c. Drug stockpiling
   d. All of the above
   e. None of the above
5. In the United States, the major factor in the cause of death is:
   a. Terrorism
   b. Smoking
   c. Accidents
   d. Cancer
   e. Heart disease
6. The Kermack–McKendrick model of epidemic assumes:
   a. Uniform mixing of individuals.
   b. Children and women are more likely to contract SARS.
   c. Public health workers are more likely to contract SARS.
   d. Diseases are either SIR or SIS.
   e. H1N1 is a persistent disease.
7. Social network models of pandemics relate spectral radius to:
   a. The number of blocking nodes
   b. The rate of spreading
   c. Normalized degree and betweenness
   d. All of the above
   e. None of the above
8. The Center for Disease Control (CDC) categorizes agents according to:
   a. Contagiousness
   b. Cost
   c. Potential to become a pandemic
   d. Impact on public health
   e. Causes of death
9. Which theory is NOT an explanation of the quick halt to SARS in 2003?
   a. Levy flights were too long.
   b. Levy flights were too short.
   c. Public health agencies were quick and effective.
   d. Air travel killed the virus.
   e. Quarantines.
10. The most (simulated) effective and practical strategy for preventing the spread of a contagious disease through the OpenFlight500 commercial air travel network is to:
    a. Harden the top 20 airports as ranked by degree and betweenness
    b. Harden the blocking nodes
    c. De-percolate routes
    d. Close down the air transportation sector
    e. Install checkpoints in all airports
11. A virulent contagion spreads without bound if:
    a. $\gamma < \Delta/\rho$.
    b. $\gamma > \Delta/\rho$.
    c. $\gamma = \Delta/\rho$.
    d. There are no vaccines.
    e. Air travel is not stopped.
12. Which of the following is NOT in the 2010 Healthcare and Public Health Sector-Specific Plan co-developed by DHHS and DHS?
    a. Eradicate pandemics
    b. Provide supplies and services during a disaster
    c. Protect healthcare workers during a disaster
    d. Mitigate risks to physical assets
    e. Mitigate risks to cyber assets
13. BioShield reportedly has:
    a. Eradicated threats from anthrax
    b. Eradicated chem/bio threats from airports
    c. Stockpiled up to $50 billion in drugs
    d. Stockpiled enough smallpox vaccine to inoculate the US population
    e. Stockpiled gas masks for all healthcare workers
14. Medical benefit payments were growing in 2013 at a rate compared to the US economy:
    a. Doubling over the next 15 years
    b. Leveling off due to the Affordable Care Act of 2010
    c. Declining until 2023
    d. Remaining constant at about 4.5% of GDP
    e. Exceeding all other government spending by 2030

15. GeoSentinel is a:
    a. CDC bio-surveillance data collection network
    b. Public Web site that processes tweets from social networks to predict outbreaks
    c. Consumer-driven Web site for tracking pandemics
    d. Division of www.Healthcare.gov
    e. Division of www.Ready.gov

14.12 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. There is little evidence that the SARS epidemic spread to other countries through in-flight contact with infected passengers. Rather, it spread because already infected passengers got off the airplane in distant cities. Propose a strategy for public health workers attempting to stop pandemics.

B. Figure 14.2 shows a monotonically increasing cost of government spending on health care in the United States. Is this evidence of the paradox of enrichment or something else. Explain why you think costs are rising with GDP.

C. Explain the underlying assumptions of the base reproduction number versus the use of spectral radius in predicting whether a contagion will become widespread.

D. Explain how big data collected from smart phones might be used in the United States to predict and prepare for outbreaks of the common flu.

E. What governs the speed of an epidemic or pandemic? Use base reproductive number, spectral radius, infectiousness, and speed of response in your answer.

REFERENCES

[1] Mayer, H. A. First Responder Readiness: A Systems Approach to Readiness Assessment Using Model-Based Vulnerability Analysis Techniques. Naval Postgraduate School, Thesis, September 2005.

[2] Tucker, J. B. and Zilinskas, R. A. The Promise and Perils of Synthetic Biology, *The New Atlantis*, 12, Spring 2006, pp. 24. Available at http://www.thenewatlantis.com/archive/12/tuckerzilinskas.htm. Accessed July 3, 2014.

[3] Kermack, W. O. and McKendrick, A. G. A Contribution to the Mathematical Theory of Epidemics, *Proceedings of the Royal Society A*, 115, 1927, pp. 700–721.

[4] Lewis, T. G. *Network Science: Theory and Applications*, Hoboken: John Wiley & Sons, Inc., 2009, pp. 511.

[5] Lewis, T. G. *Bak's Sand Pile*, 2nd ed, Monterey: AgilePress, 2011, pp. 377.

[6] Hu, Y., Luo, D., Xu, X., Han, Z., and Di, Z. Effects of Levy Flights Mobility Pattern on Epidemic Spreading under Limited Energy Constraint, arXiv:1002.1332v1 [physics.soc_ph], February 5, 2010. Available at http://www.researchgate.net/publication/45899671_Effects_of_Levy_Flights_Mobility_Pattern_on_Epidemic_Spreading_under_Limited_Energy_Constraint

[7] Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., and Buckee, C. O. Heterogeneous Mobile Phone Ownership and Usage Patterns in Kenya, *PLoS ONE*, 7, 4, 2012, e35319. doi:https://doi.org/10.1371/journal.pone.0035319.

[8] Minino, A. M., Arias, E., Kochanek, K. D., Murphy, S. L., and Smith, B. L. Deaths: Final Data for 2000, *National Vital Statistics Reports*, 50, 15, 2002, pp. 1–20.

[9] Mokdad, A. H., Marks, J. S., Stroup, D. F., and Gerberding, J. L.. Actual Causes of Death in the United States, 2000, *JAMA*, 291, 10, 2004, pp. 1238–1246.

[10] Ghosh, J. K. and Valtorta, M. Probabilistic Bayesian Network Model Building of Heart Disease. Technical Report TR9911 USCEAST-CSTR-IY99-11. Department of Computer Science, University of South Carolina, November 30, 1999.

15

TRANSPORTATION

The transportation CIKR consists of highways, bridges, tunnels, traffic control systems; trucks, buses, and other commercial vehicles; railroads, rail rolling stock, rail yards, rail SCADA and signaling systems, rail bridges, and tunnels; oil and gas pipelines safety; intermodal facilities such as seaports, intracostal ports and waterways, marine terminals, marine vessels, containers, barges, locks, and dams; air carrier airports, general aviation airports, air navigation and traffic control systems, airfreight and package express systems, and passenger and cargo aircraft; and commuter rail and public transit systems, passenger ferries, Amtrak, and intermodal passenger terminals.

The transportation sector is transitioning from disparate land, air, rail, and merchant marine subsystems focused mainly on efficiency and profitability to an integrated intermodal network focused more on safety, security, and regulation. Complexity comes from integration of intermodal services across roads, commercial air, rail, and shipping—both for passengers and freight. Complexity also comes from increasing regulation in the form of safety rules, mileage goals for passenger and freight vehicles, modernization of air travel, and promotion of high-speed intercity rail.

In the United States this transition is being led by the Department of Transportation (DOT) as the sector-specific agency (SSA) for transportation and Department of Homeland Security (DHS) under PPD-21 (2013). Established by an act of Congress on October 15, 1966, the mission of the department is to "Serve the United States by ensuring a fast, safe, efficient, accessible and convenient transportation system that meets our vital national interests and enhances the quality of life of the American people, today and into the future."

In addition to fast, safe, and efficient transportation, the transportation challenge facing the United States is to sustain roads, bridges, railways, commercial airports and navigation systems, and intermodal ports, as costly maintenance, repair, and technological change sweeps over the sector. Without large investments going forward, transportation is threatened by imminent decay and disrepair. And without modernization to accommodate automation of air traffic, electric cars that drive themselves, and heightened pressure to "plug in" to the global supply chain, the US transportation system will become a drag on the economy.

The following topics survey just a fraction of the factors affecting the transportation CIKR sector:

- *Transportation is vast*: The transportation sector is a vast network of networks including highways, railroads, commercial air travel, gas and oil pipelines, and numerous agencies in the public and private sector. Governmental regulation is split between the DOT, which is focused on safety and sustainability, and the DHS's Transportation Security Administration (TSA), which is focused on security and counterterrorism.
- *Transportation is in transition*: The transportation sector is in a long-term transition period due to socioeconomic and political forces: deregulation since 1978, shifts in energy policy stemming from concerns about the environment, global intermodal networks

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

underpinning the nation's vital supply chains, and rapid technological change are reshaping transportation.

- *Rising costs*: Like so many critical infrastructures, the cost of maintaining and modernizing the *National Highway System* (NHS) is rising faster than the country's gross domestic product (GDP). Highways may be subject to the tragedy of the commons, especially if gasoline taxes continue to decline due to improvements in gasoline engine efficiency and the transition to electric cars. A political focus on mass transit and intercity rail exacerbates the problem, raising questions regarding the sustainability of highways.
- *Interstate nation-builder*: The Dwight D. Eisenhower National System of Interstate and Defense Highways—Interstate Highway System—celebrated its 50th anniversary in 2006 after building out over 47,000 miles of roadway at an estimated cost of $425 billion. Conceived in the 1930s and 1940s, this super highway is a small fraction of the 3.9 million miles of highway crisscrossing the nation but has had a significant impact on the US economy and way of life.
- *Redirection/Transition*: The Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991 began the transition from the traditional point-to-point strategy to an integrated intermodal transportation strategy by redirecting funds to other modes—mass transit, intercity rail, and supply chain connections to ports and railways. This legislation marks the beginning of transition in this sector.
- *Road resiliency*: The 160,000-mile NHS of roads and bridges is extremely resilient. The consequences to the economy from a 1-year loss of critical bridges across the Mississippi River and tunnels connecting East and West are minor, because of the low spectral radius and minimal betweenness of the system. Highways support most of the movement of cargo as well as people and have been shown to be extremely robust and resilient by a number of studies. Analysis of the Interstate Highway System network has low spectral radius (4.3), high node and link redundancy, and 64 blocking nodes (critical connector cities).
- *Disruptive railroads*: The rapid rise of railroads between 1830 and 1857 surprisingly mirrored the rapid rise of the Internet between 1970 and the dotcom crash in 2000. More importantly, the railroads were the first technology bubble with major side effects on society. Railroads were the stimulus for big corporations, which led to the first government regulation of the private sector—the Interstate Commerce Act of 1887. Railroad technology was the first "hi-tech" to obey the technology diffusion curve that all hi-tech companies track today. In addition, railroad changed modern life in profound ways similar to the impact on society of the Internet.
- *Competitive exclusion*: The railroads—and hi-tech companies like Microsoft—prove once again the power of Gause's competitive exclusion principle and the economics of increasing returns. At one point in history, Commodore Vanderbilt owned enough of the railroad network to control the industry, just as Microsoft owned enough of the software business to control the software industry. However, railroad monopoly in the 1880s led to government intervention in the form of regulation by the Interstate Commerce Commission (ICC), while the software monopoly held by Microsoft in 1998 did not. This pattern of regulation followed by deregulation shapes most critical infrastructure systems today. Thus, the rail transportation system established a model used by Congress to regulate—and eventually deregulate—entire industries.
- *Commuter rail*: While there is less than 8000 miles of commuter rail in the United States, it is increasingly important as a people move in dense metropolitan areas. Commuter rail is characterized by not only low spectral radius (resilience against cascades) but also very low robustness (almost all nodes and links are critical). Moreover, the number of terrorist attacks on passenger rail far exceeds the number of terrorist attacks on commercial airliners: 181 passenger rail attacks in the period of 1998–2003 versus 8 commercial airliner attacks in the period 1986–2012. Commuter railway stations are easy targets for terrorists seeking to kill large numbers of people.
- *Air travel follows rail*: Commercial air travel in the United States followed the same technology diffusion model established by the railroads—rapid build-out followed by saturation, extreme competition, and lackluster sustainability. Accordingly, the commercial airline industry has cycled through the regulate–deregulate transition pioneered by railroads. Historically, regulation has been enacted to reduce competition, and deregulation has been used to increase regulation. In the case of the airlines, deregulation restored competition with the passage of the *Airline Deregulation Act* of 1978, after decades of tight regulation of routes and fares. Rather than preventing a monopoly or commercial air travel, regulation prevented its sustainability. Rather than leading to a monopoly, deregulation appears to be promoting competition.
- *Airline network structure*: Today's hub-and-spoke network structure of domestic (and international) airports and routes is a consequence of deregulation. The hub-and-spoke network is efficient and reduces operating costs, but it also leads to low resiliency and low robustness in terms of betweenness and blocking nodes. The mean connectivity of the US domestic hub-and-spoke network is only 8.33, but its spectral radius is 50.8, and

2% (64) of airports are blocking nodes. It is prone to complex catastrophic collapse—its critical vulnerability is only 4.6%.

- *Airline safety*: The commercial air travel segment of transportation has an enviable record of safety. Safety is regulated by the National Transportation Safety Board (NTSB), which was purposely established outside of DOT so that if could perform its functions totally separate and independent from any other agency. The risk of dying in a commercial airliner is so small that it may not be statistically valid—odds are literally less than 1 in a million.
- *Air travel security*: The TSA (an agency within the DHS) is responsible for air travel security. As such, it is responsible for highly controversial security measures ranging from passenger pat-downs at airports to surveillance of passengers through electronic means. The Terrorist Screening Center (TSC) implements the Secure Flight information system that stores passenger names, gender, birth date, and frequent flyer information on passengers. Critics claim that TSC violates the 1974 Privacy Act, which prevents the government from abusing the privacy of its citizens.
- *Airport games*: TSA terrorist countermeasures include game theoretic tactical software in addition to surveillance of passengers to protect passengers against terrorism. GUARDS is a game theoretic software program that maximizes allocation of limited defensive resources while minimizing risk within airports. It simulates a two-person competitive game—attacker versus defender—in an airport setting. Defenders allocate limited resources to protect vital targets against random terrorist attacks on the same targets. GUARDS "discovers" the best mixed strategy of randomized countermeasures to apply at each airport.
- *Bayesian networks*: An alternative to GUARDS and game theory is the application of Bayesian belief networks to anticipating and fending off terrorist attacks. Bayesian networks are models of cause and effect, whereby the probability of an attack increases or decreases as evidence is gathered and plugged into the model.

15.1 TRANSPORTATION UNDER TRANSFORMATION

The US transportation safety and security structure is shown in Figure 15.1. The transportation sector consists of highways, railroads and commuter trains, commercial and general aviation, and gas and oil pipelines (see Fig. 15.2). Roads and railways have been recognized as nation-builders and economic accelerators from the earliest days of the nation. In the 1840–1940 era, railroads bound east and west together and brought prosperity to every region they connected. Similarly, a century later, commercial air travel linked nations together to form one global commons rivaling even the Internet.[1] In the twenty-first century transportation will play an even greater role in intermodal movement of people and goods around the globe.

Intermodal transportation means the flow of people and goods using different modes—automobile passengers change mode when they connect to an airport that takes them to another country where they use buses, automobiles, and trains to reach their ultimate destination. Intermodal freight transportation means the movement of freight by truck to a ship in a Chinese port to a consumer in the United States via ships, airplanes, trucks, and rail. Intermodal transportation has become an industrial commons of different modes of transportation that work seamlessly together. Note that gas and oil pipelines are part of the transportation sector; for historical reasons (see Chapter 12).

The rapid rise of intermodal transportation connecting people and products from around the world has improved the living standard of everyone on the planet due to the economics of *comparative advantage*. Goods and services inexpensively flow with relative ease from any place on the globe to anywhere else. Globalization increases the standard of living of trading nations. But it has also heightened the threat of terrorism, increased the spread of diseases, and promoted economic dislocations. Terrorists can easily move around the globe, diseased travelers carry infections such as SARS with them as they travel, and jobs go to regions of the world where labor is cheapest. Like the Internet, intermodal transportation is reshaping the world along supply chain corridors and major ports of call.

As Figure 15.1 shows, the reaction of many governments to rapid change in the transportation commons has been to regulate it for a while and then deregulate it in an effort to strengthen national infrastructure. In the United States regulation is managed by the US DOT, which regulates safety and serves as the SSA for transportation security for the DHS. DOT is a relatively large organization, consisting of agencies to oversee—and fund—roads, railroads, airways, and gas and oil pipelines. (Its budget in 2012 was $70 billion and rising.)

DOT was created in 1967 to pull together a number of existing agencies such as the Federal Aviation Administration (FAA) and NTSB. (The NTSB was later separated from the DOT.) The US Merchant Marine Academy located at Kings Point, New York, was dedicated in 1943 to train merchant marines. It was an outgrowth of the Shipping Act of 1916

[1]There were approximately 42,000 registered autonomous systems and 375,000 peering relations (links) in the 2012 Internet and approximately 45,000 airports with 62,000 routes (links) in the OpenFlight network of airports and routes.

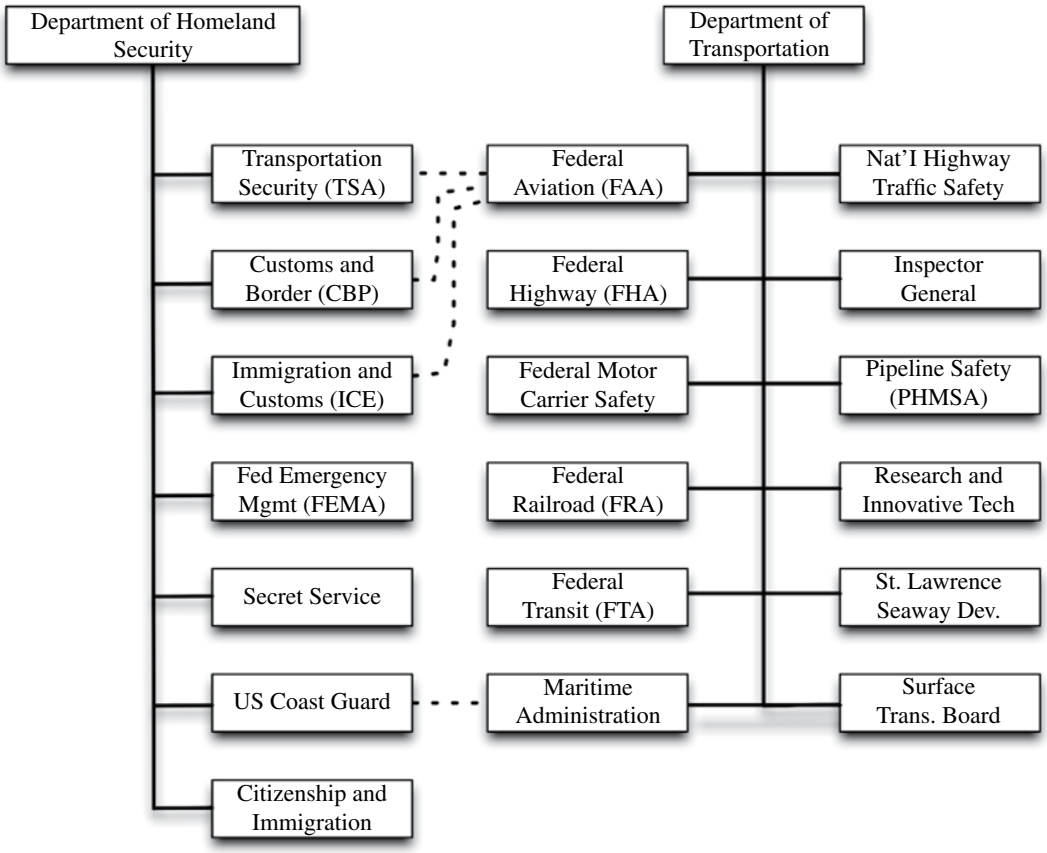


FIGURE 15.1 The Department of Transportation is the sector-specific agency for the Department of Homeland Security.

and the Merchant Marine Act of 1936. President Truman's Reorganization Plan No. 1 established Merchant Marine Safety Administration (MARAD) in 1950, which was transferred to DOT in 1981. Similarly, the St. Lawrence Seaway Development Corporation, a wholly government-owned corporation for running two locks along the St. Lawrence Seaway, along with Canada, became the responsibility of DOT.

Initially, DOT's mission was focused on safety—seat belts and airbags for cars and airline safety. Then the mission expanded to include environmental focus—Corporate Average Fuel Economy (CAFÉ) standards that aim to increase automobile mileage and reduce pollution. Today it has ambitious goals of promoting intercity high-speed rail, maintenance of highways and bridges, and next-generation global navigation systems for commercial airliners. Its expanded goals may outstrip funding, which is principally based on fuel and commercial transportation taxes and the will of Congress.

The FAA was created in 1958 to regulate aviation safety. The Federal Highway Administration (FHWA) and Federal Railroad Administration (FRA) were created in 1966 to manage the *Highway Trust Fund* —the source of funding for the Interstate Highway—and NTSB was created in 1967 for highway safety. The Federal Motor Carrier Safety Administration (FMCSA) was spun off of FHWA in 2000 pursuant to the Motor Carrier Safety Improvement Act of 1999. The National Highway Traffic Safety Administration (NHTSA) created in 1970 regulates CAFÉ standards, seat-belt, child seat, and airbag safety, and anti-theft standards. The Federal Transit Authority (FTA) manages funding of public transportation systems such as commuter light rail.

The Mineta Research and Special Programs Improvement Act of 2004 integrated the Offices of Pipeline Safety (OPS) and Hazardous Materials Safety (OHMS) under one roof—the Pipeline and Hazardous Materials Safety Administration (PHMSA). The Surface Transportation Board was created in 1995 to regulate non-gas and oil pipelines. Research and Innovative Technology Administration (RITA) created in 2004 coordinates research and education programs in inter-modal technologies. For example, new digital technologies are being applied to tagging and tracking shipping containers around the world.

Today, almost all aspects of transportation are regulated or partially regulated by some government agency. And as regulation increases and infrastructure ages, costs also increase. Figure 15.3 shows the exponential growth of just one DOT program—revenues devoted to the Interstate Highway

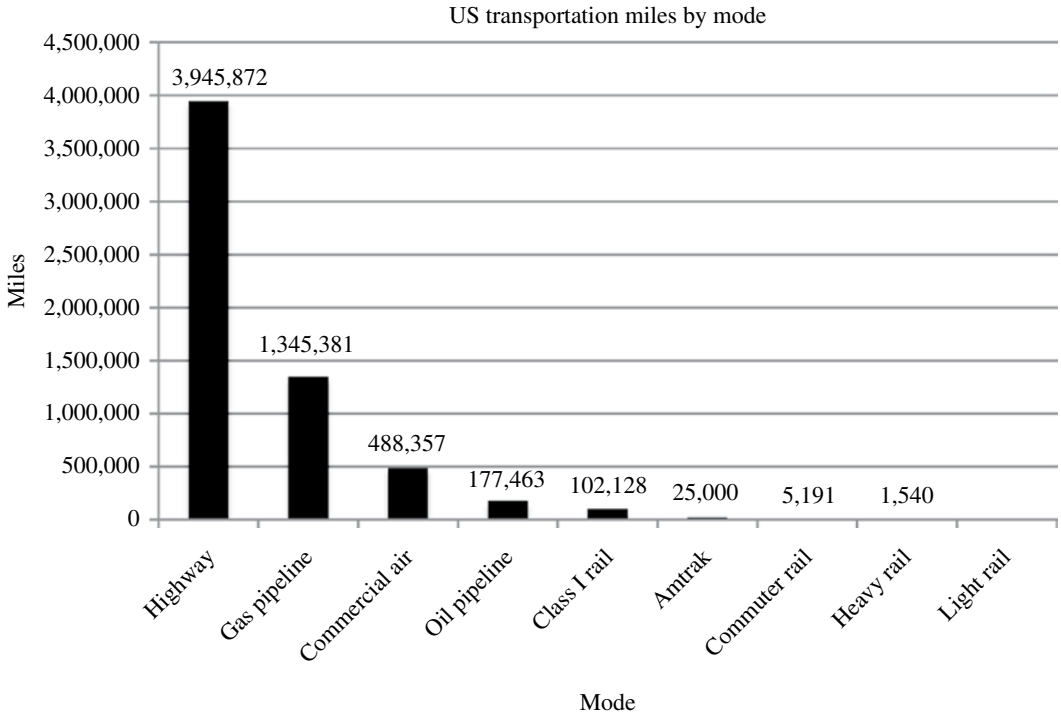


FIGURE 15.2 The intermodal transportation system of the United States consists of highways, gas and oil pipelines, commercial and general aviation, cargo and passenger rail, commuter rail, and heavy and light rail.

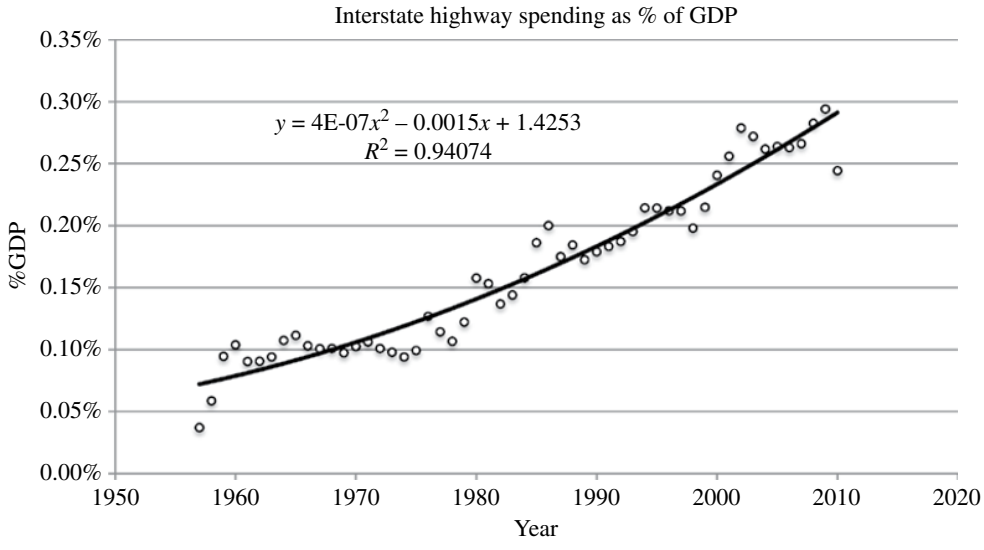


FIGURE 15.3 The cost of building and maintaining the 50+ year-old Interstate Highway System is growing much faster than gross domestic product (GDP).

System. Maintenance and repairs eventually overtake expansion in accordance with the paradox of enrichment. Adding to this complexity is the impending reduction in revenues from gasoline taxes, as electric cars and natural gas trucks replace gasoline-powered vehicles. (90% of the Highway Trust Fund comes from federal gasoline taxes or Congress, and 10% comes from the states, typically through a local tax on fuels.)

So the number one threat to the transportation CIKR is lack of sustainability. Are transportation systems subject to the tragedy of the commons? Have some portions of the intermodal transportation network grown beyond their *carrying capacity*? Have new technologies rendered legacy transportation systems of the twentieth century inadequate for the age of the Internet? Highways and rail may become

irrelevant or inadequate in an age of driverless cars and railroad and flying drones.

15.2 THE ROAD TO PROSPERITY

By all accounts, the *Dwight D. Eisenhower National System of Interstate and Defense Highways*—known as the *Interstate Highway System*—has been a phenomenal success. Like the railroads a century earlier, the Interstate was a nation-builder. According to Wikipedia[2]:

> The system is named for President Dwight D. Eisenhower, who championed its formation. Construction was authorized by the Federal Aid Highway Act of 1956, and the original portion was completed 35 years later. The network has since been extended, and as of 2010, it had a total length of 47,182 miles … making it the world's second longest after China's. As of 2010, about one-quarter of all vehicle miles driven in the country use the Interstate system. The cost of construction has been estimated at $425 billion in 2006 dollars.

Since its inception in 1956, the Interstate Highway System and the larger NHS have undergone a number of funding and regulatory changes. ISTEA of 1991 was the first post-Interstate Highway act to redirect national strategy to focus on intermodal integration and mass transit systems. It was followed by a series of TEAs—Transportation Equity Act for the Twenty-First Century (TEA-21) in 1998 and Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU) in 2005 and Moving Ahead for Progress in the Twenty-First Century (MAP-21) Act in 2012. Each of these laws moved highway infrastructure a step closer to a highly integrated, intermodal, and regulated system.

15.2.1 Economic Impact

A study conducted for DOT in 2006—the fiftieth anniversary of the Federal Aid to Highways Act of 1956—concluded the following [1]:

- Interstate Highway investments have lowered production and distribution costs in virtually every industry sector. On average, US industries saved an average of 24 cents annually for each dollar invested in non-freeway `roads.
- Interstate Highway investments contributed 31% to annual productivity growth in the 1950s, 25% in the 1960s, and 7% in the 1980s. The declining contribution to productivity growth is reflected in the exponential increase in cost shown in Figure 15.3. There are two major reasons for this decline: (1) funds derived from gasoline taxes and congressional appropriations to the Highway Trust Fund have been diverted to non-highway projects such as mass transit, and (2) maintenance costs rise as the system ages. For example, maintenance and repair bills for 2002 were $24 billion, but only $10 billion was available.

Although travel by highway exceeded 3 trillion passenger miles in 2012, productivity improvements are largely realized in the movement of freight. Intermodal transportation—trucks and rail, mainly—moved 21 billion tons of freight worth $15 billion in 2006. Sixty percent of this was moved over highways, from 17 international gateways, across 600,000 bridges, through more than 50 major tunnels on the way to and from 114 major economic regions called *centroids*. This complex CIKR makes up a road network commons known of as the National Highway System.

15.2.2 The National Highway System (NHS)

The Interstate system is a small part (47,000 miles) of the 160,000-mile NHS consisting of 5 major components:

1. The Interstate Highway System of freeways and turnpikes.
2. The Principal Arterial connectors that link major roads and highways to ports, airports, public transportation facilities, and other intermodal transportation facilities.
3. The Strategic Highway Network (STRAHNET) that provides access to and from major military bases and storage depots.
4. The Strategic Highway Connectors that link major military installations to the STRAHNET.
5. The other Intermodal Connectors linking intermodal facilities to the other four subsystems making up the NHS.

As it turns out, this is the most resilient and robust CIKR studied in this book. A thorough study of the NHS by a group of researchers in 2010 concluded that the collapse of critical bridges linking the Western half of the United States to the Eastern half would have little impact on the economic well-being of the Unites States [2]. They studied the impact of 1-year closures of two and four bridges across the Mississippi and closure of the Eisenhower Memorial Tunnel along I-70 west of Denver, Colorado. The two-bridge scenario examined the impact of rerouting freight traffic around the I-55 Memphis–Arkansas crossing; the four-bridge scenario examined the impact from closure of the I-10 Horace–Wilkinson, I-74 Iowa–Illinois, US-67 Rock Island Centennial, and I-280 Iowa–Illinois crossings.

[2]https://en.wikipedia.org/wiki/Interstate_Highway_System

Re-routing through shortest paths increases shipping costs—typically amounting to $0.90/mile, which in turn increases commodity prices. It also increases travel time and fuel costs. But does it significantly impact the national economy? The researchers found consequences highest for the two-bridge scenario followed by the four-bridge and tunnel scenarios. But the consequences from all scenarios are modest because the NHS contains redundancy.

Consequences were measured in Passenger Car Equivalents times hours—PCE*hours. The two-bridge scenario cost 3061 million; four-bridge cost 674 million; and tunnel closure cost 576 million PCE*hours over a 1-year period. While these are sizeable losses, they pale in comparison with the 3 trillion PCE*hour total for the nation. Therefore, the researchers concluded that the NHS is resilient and robust even when multiple bridges and key tunnels collapse. This is in line with the relatively low spectral radius of the NHS.

15.2.3 The Interstate Highway Network Is Resilient

Similar results can be found for the 47,000-mile Interstate system even though it has far fewer routes. Figure 15.4 shows the freeway network formed by 281 cities (nodes) and 384 freeway segments (links) connecting major regions of the United States. This network is a small subset of the NHS, but even this small subset is robust and resilient. Moreover, link and node robustness are 27 and 77%, respectively. Over 100 links can be removed without disconnecting the network. Only 23%—(0.23)(281) = 64 cities—are blocking nodes. There are many alternative routes connecting East and West and North and South. The topology of the Interstate Highway Network approximates a random network, which is the most robust and resilient class of networks known.

The Interstate network has a relatively large diameter—it takes 34 hops from city to city to travel to and from the most distant cities. Chicago is the hub, and the cities with the highest betweenness are in the middle, as expected. The top 10 cities in terms of betweenness are:

Indianapolis (5961 paths)
Oklahoma City
Louisville
Kansas City
Birmingham
Albert Lea
Erie
Atlanta
Champaign
Echo

The average betweenness across all cities is relatively high (23%), which indicates load balancing—traffic is spread across all cities more or less evenly. Taken altogether, the Interstate Highway Network is nearly impossible to severely damage. It would require a complex catastrophe more consequential than a major earthquake, extreme superstorm, megahurricane, or series of coordinated terrorist attacks to render it useless. Indeed, the Interstate Highway System has survived complex catastrophes such as the 2005 Hurricane Katrina, Superstorm Sandy in 2012, the I-35 Minneapolis bridge collapse in 2007, and the I-40 bridge collapse in 2002.

15.2.4 The NHS Is Safer

The highways of the United States are safer today than ever before, due to a number of technological advances. Roads are safer, cars are equipped with safety devices like seat belts and adaptive cruise control, and traffic control is more sophisticated. Soon automobiles will be automated, which will reduce human error even further.

Deaths due to automobile collisions are still too high, but from a statistical point of view, risk is declining. In 1921, there were 24 deaths per 100 million miles traveled. By 2017 that number declined to 1.2. Deaths per capita peaked in 1937 to 29.4 and dropped unevenly to 11.4 in 2017.[3] As both population and miles driven increased, deaths on the US highways decreased.

Unlike most of the catastrophic failures studied in this book, transportation casualties do not obey long-tailed power laws. They are not normal accidents according to NAT. They tend to obey the normal distribution, although this is difficult to show because the statistical data is nonstationary. That is, their average values change over time as described above.

Ignoring the nonstationary nature of transportation accidents and fatalities for a moment, the fact that transportation deaths do not obey long-tailed statistics says something about the nature of transportation versus other CIKR. Recall that the underlying mechanism that produces a normal distribution is randomness, while the underlying mechanism that produces long-tailed distributions is connectedness, or conditional probability. Normal accident theory describes how complex systems fail because they are made of coupled parts. This coupling—connectedness—imposes dependencies. We model these dependencies as conditional probabilities rather than independent probabilities. The result is long-tailed distributions.

15.3 RAIL

In many ways, the emergence of railroads from 1830 to 1857 parallels a similar rise of the Internet from 1970 to 2000. Both technology developments took 25–30 years to mature,

[3]https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year

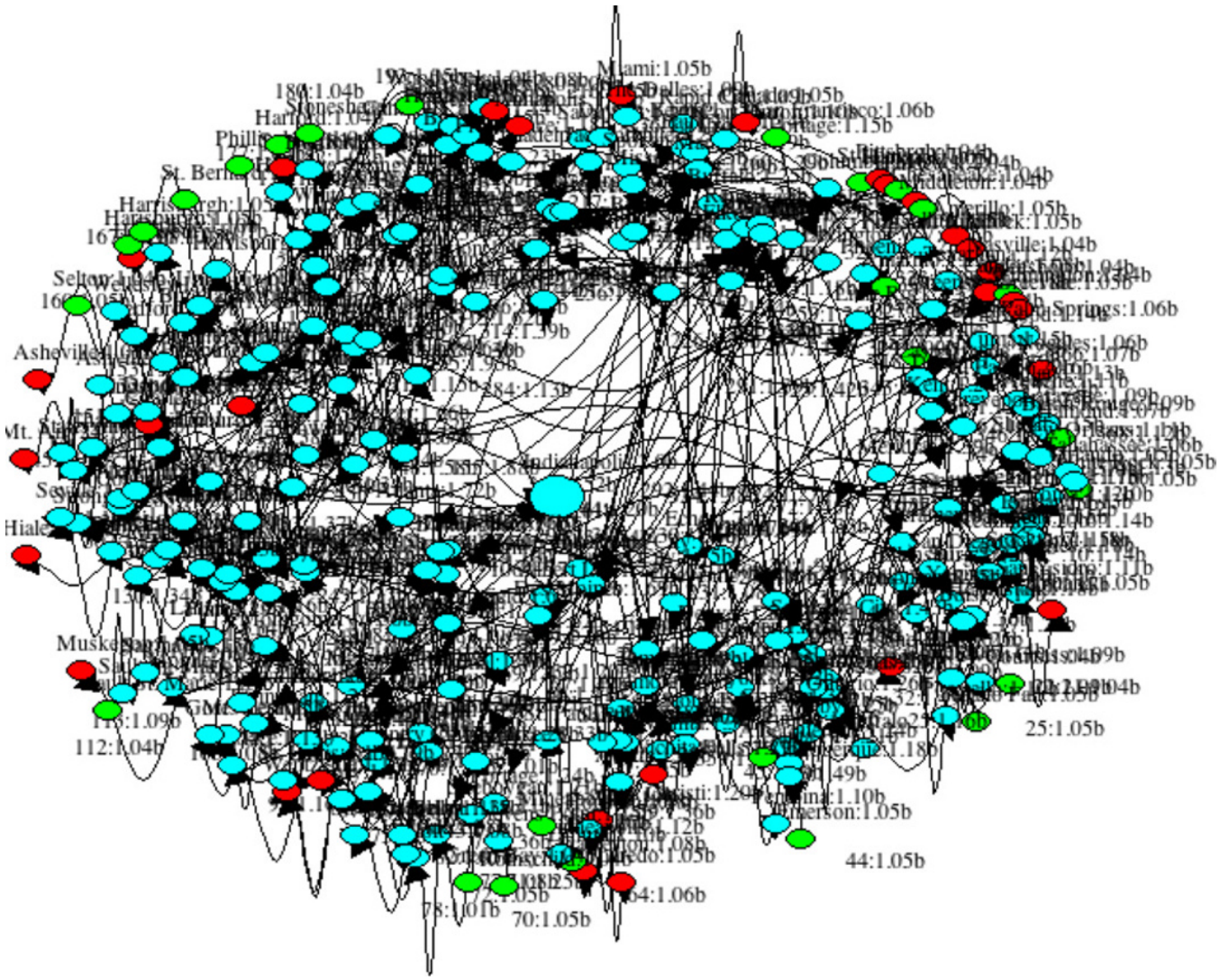


FIGURE 15.4 The Interstate Highway System forms a transportation network containing 384 connecting roads (links) and 281 cities (nodes), with spectral radius equal to 4.3.

and both started out to solve one problem and ended up solving another. George Stephenson combined James Watt's steam engine with coal wagon technology to construct the steam locomotive. His objective was to extract coal from mines near Leeds, England, in 1811, but his steam locomotive found a much broader market. He rapidly improved on his initial design to stay ahead of the competition, and by the late 1820s his steam-powered locomotives established a new mode of general transportation. Stevenson's locomotive, the *Rocket*, traveled 12 miles in 53 min—becoming the centerpiece of the Liverpool and Manchester Railway in 1830. Twenty-five years later, Stevenson's invention formed the basis of an entire industry that affected every aspect of modern life.

Similarly, the TCP/IP protocol evolved out of early work on computer networking, beginning in the late 1960s and early 1970s. Robert Taylor was interested in solving the "terminal problem" (how to connect several different mainframe computers to one user terminal) so that fewer terminals would clutter his desk. This led to the first ARPANet spanning a small geographical area. In 1973–1974, Vinton Cerf and Robert Kahn developed Transmission Control Protocol (TCP), and DARPA required its use in 1976, thus standardizing network communication. The solution to Taylor's terminal problem led to the first version of the Internet. Instead of connecting a few terminals to a few mainframes, TCP/IP connects nearly every type of communicating machine to every other machine in the world today. The Internet formed the basis of an entire industry and affected every aspect of modern life.

Railroads were one of the first technology deployments to obey the *technology diffusion* curve shown in Figure 15.5

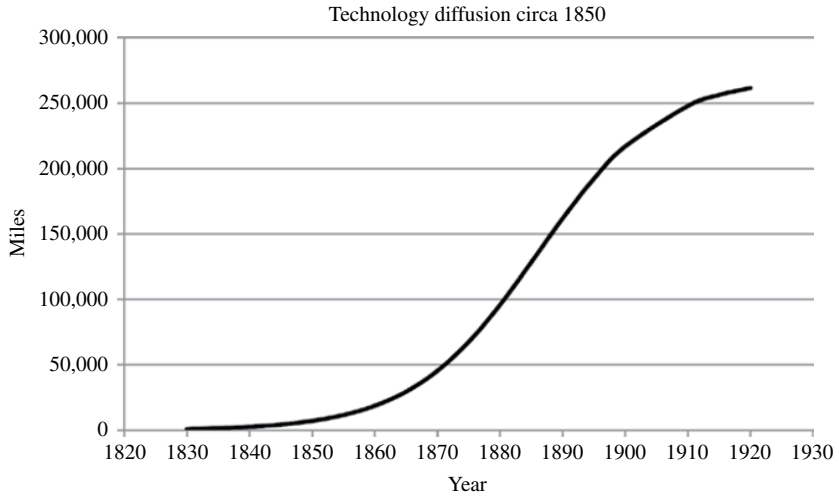


FIGURE 15.5 Railroad technology was one of the earliest examples of technology diffusion, which obeys an S-shaped logistics function over time.

and so familiar to modern infrastructure adoption today. Diffusion starts small, gathers momentum and spreads exponentially for a short period of time, and then flattens out as saturation takes over. The spread of automobile, radio, TV, telephone, and Internet technology all obey this curve. It signals acceptance of a breakthrough infrastructure and often causes major societal changes.

The US economy and railroads rapidly expanded from 1852 to 1857, ending in the Panic of 1857. Similarly, the Internet rapidly expanded in 1992–2000, ending in the *dot-com bubble* crash. Both recovered and continued their diffusion until reaching a point of market saturation. By 1920, most of the modern railroad infrastructure was in place, and by the mid-twenty-first century, the global Internet will be in place. The railroads reached saturation within landlocked regions such as North America, Europe, and Asia. The Internet will reach saturation when every person on the planet is connected.

Competitive exclusion and *increasing returns* impacted the railroad industry in its early years, just as competitive exclusion and increasing returns impacted the twentieth-century hi-tech industry. Competitive exclusion explains Microsoft's rise to a monopoly position in the 25-year period 1975–2000.[4] The computer industry made William Gates the wealthiest person in America, and similarly, railroads made Cornelius Vanderbilt the wealthiest person in America in the 1880s. Vanderbilt sold all of his steamships in 1864 to focus on railroads. Within 15–20 years he dominated the railroad sector. Vanderbilt's legacy includes Grand Central Station, which he started constructing in 1885. Railroads, pipelines, and commercial airline infrastructure industries are examples of Gause's law in action.

[4]Microsoft was deemed a monopoly in violation of the 1890 Sherman Antitrust Act in 1998.

15.3.1 Birth of Regulation

But there are several important differences between technology diffusion in the first century of industrialization and today's information age. The rise of rails corresponded with the rise of big business and precipitated the Interstate Commerce Act of 1887, which marks the beginning of a century-long period of regulation of business by government. The ICC was set up to monitor big business and settle disputes among competitors. The Sherman Antitrust Act of 1890 soon followed, giving the government a big stick to regulate big business. It broke up trusts and prevented monopoly power in one business from being used to achieve a monopoly in another business. This law has had a profound impact on shaping critical infrastructure, as discussed in previous chapters.

According to a study done for the World Bank, "The first nationwide regulation of transportation in the U.S. was intervention in railways: interestingly, it came about because of a belief that there was too much competition. In the 1830 to 1880 period, railways had been over-built in many areas of the country—especially the Northeast—mainly because of financial speculation in the creation of railway companies" [3]. Government intervention in privately owned and operated businesses began with the Interstate Commerce Act of 1887, followed by a series of increasingly restrictive legislation:

- Elkins Act of 1903: Outlawed rebates and kickbacks to shippers (customers).
- Hepburn Act of 1906: Permitted the ICC to put a ceiling on shipping rates and extended regulation to pipelines.
- Mann–Elkins Act of 1910: Further restricted prices and allowed shippers to designate the route taken by a shipment.

- World War I: The federal government took over the operation of the railroads in a move reminiscent of corporate bailouts following the economic meltdown of 2008–2009.
- Transportation Act of 1920: Allowed ICC to set shipping rates so that railroads could make a 6% return, redistributed profits in excess of 6% to weaker railroads, and set minimum rates.
- Emergency Transportation Act of 1933: Propped up railroads but increased federal intervention in management of privately held railroads.
- Motor Carrier Act of 1935: Added trucking to the regulatory oversight of the federal government, reduce competition, and stabilize rates.
- Transportation Act of 1940: Added water carriers to the list of regulated transportation companies.
- Reed–Bulwinkle Act of 1948: Legalized rate-setting cartels under ICC control.

By 1970, the rail industry was losing $300 million/year on passenger service, and its market share of freight dropped from 75 to 37% by 1988—9.6% of total freight revenues. Penn Central Railroad entered bankruptcy 3 years after a gigantic merger of large Eastern railroads. This led to nationalization of Penn Central, which changed its name to Conrail. This action by Congress marked a turning point. Starting in the 1980s, Congress began to deregulate transportation:

- The Staggers Act of 1980: Relaxed price setting by the ICC and allowed carriers more leeway in running their businesses.
- Motor Carrier Act of 1980: Deregulated entry into trucking industry to increase competition, deregulated rates, and replaced regulatory controls with antitrust restrictions.

The history of transportation regulation followed by deregulation mirrors the change in policy over the past 150 years. In contrast to the rise of government regulation and big business in the nineteenth century, the later part of the twentieth century was characterized by just the opposite—a megatrend toward deregulation. Perhaps the most startling example of this turnabout is the case of Microsoft versus the US Department of Justice. Even though Microsoft was found guilty of violating Section 2b of the Sherman Antitrust Act in 1998, it was not broken up or regulated like earlier big businesses. Instead, Microsoft was modestly fined and allowed to continue to operate as a nonregulated entity. Even the highly regulated energy and telecommunications sectors were partially deregulated through the 1992 EPACT and 1996 Telecommunications Act. As a further sign of the times, the Internet operates as an unregulated infrastructure unfettered by regulation in the United States and other liberal democracies today.

Regulation, followed by deregulation and changes in attitude toward big business, has had a profound impact on infrastructure systems. In the early part of the twentieth century, government shaped vertical monopolies and limited the consequences of Gause's competitive exclusion principle. Infrastructure systems tended to be vertically integrated monopolies or oversubscribed oligopolies. But subsequent deregulation produced a tragedy of the commons in energy and communications sectors. In rare cases, such as the Internet, both extremes were avoided. Today, the Internet has not suffered the tragedy of the commons, but there are early signs that Gause's law is actively shaping the Internet, as discussed in Chapter 5.

Congress has repeatedly applied transportation's regulate–deregulate model to other CIKR sectors. Three distinct phases emerge from the regulate–deregulate model:

1. Intervention: Government enacts a series of regulatory restrictions and laws to limit the power of a monopoly or preserve an industry.
2. In extreme cases, government grants a corporation *natural monopoly* status to gain efficiencies and universal access or preserve a critical infrastructure.
3. Eventually, the natural monopoly or marginal industry becomes obsolete, inefficient, or insolvent, so the government deregulates the industry and opens it up to a new generation competitors—typically with incentives for the introduction of advanced technology.

The last phase is especially interesting because it has been used to deregulate transportation, energy, and communications. Deregulation takes the form of commoditization of the industrial commons underlying the sector—pipelines, in the case of energy; telephone lines, in the case of communications; and rails and routes, in the case of rail and air transportation. By removing the infrastructure from monopoly protection or insolvency, the government stimulates a new round of invention and innovation.

15.3.2 Freight Trains

The freight rail industry is a mature infrastructure dominated by the private sector but heavily shaped by government regulation. It has evolved into a complex system of transportation modes spanning the globe. Freight rail is one major cog in the US intermodal supply chain stretching from major ports like Los Angeles–Long Beach and New York–Elizabeth to Chicago, where global products are distributed throughout the rest of the nation. East, West, North, and Southbound railroads and interstate highways radiate out from Chicago, making the city a major transportation hub.

A handful of freight rail corridors are critical to this national supply chain. One extends from Chicago to Seattle–Tacoma; another to Los Angeles–Long Beach and southward to El Paso, Dallas, and Texas, as part of the NAFTA trade route (North American Free Trade Agreement between the United States, Canada, and Mexico). East and Southbound rails connect Chicago to the Gulf Coast, Atlanta, and the populous Northeastern United States. There are less than a dozen critical links in the entire freight rail infrastructure serving the United States.

15.3.3 Passenger Rail

Passenger rail security is the focus of homeland security and transportation, because of a history of terrorist attacks on passenger rail. Terrorists accounted for 431 passenger rail fatalities in 181 attacks in 1998–2003, according to an RAND report published in 2004 [4]. Moreover, deaths due to attacks on trains exceeded fatalities from airplane incidents—69 airplane incidents resulted in 33 deaths—during the same time period. But bus and mass transit attacks and accidents have caused more deaths than trains.

The DHS warns of the dangers, "Passenger trains and stations are especially attractive terrorist targets because of the large number of people in a concentrated area. A terrorist attack against freight rail would require more complex planning, timing, and execution to cause high casualties or costly economic damage" [5]. The focus of attention, therefore, is on passenger rail.

The most likely terrorist weapon in an attack on passenger rail is an improvised explosive device (IED) hidden in a backpack. Four terrorists attacked passengers on the London underground train in July 2005 by detonating explosive hidden in backpacks. A suicide bomber attacked an entrance to the Moscow Metro in August 2004, and Islamic extremists used 13 remotely detonated IEDs on four commuter trains in Madrid in March 2004. IEDs are easy to construct and conceal in carry-on packages and backpacks.

Fatalities from terrorist attacks on passenger rail have been low—mostly single-digit numbers—rarely exceeding dozens of injuries and deaths. The fractal dimension of fatalities due to terrorism is 1.06, which is marginally low risk [6]. Seventy-two percent involved no fatalities, and 58% involved no injuries. The principal consequence of a rail system attack is fear, not death.

15.3.4 Commuter Rail Resiliency

Light rail commuter trains serving metropolitan areas are not particularly resilient against bombings, because they lack redundancy. Table 15.1 summarizes structural information for Amtrak and four other metropolitan commuter rail networks. Average values of spectral radius for Southeastern Pennsylvania Transportation Authority (SEPTA), Los Angeles Metro, Bay Area Rapid Transit (BART), Boston Metropolitan Transit Authority (MTA), and Amtrak indicate very low self-organization and node and link robustness. Therefore, they are resilient against cascade failures but vulnerable to node (stations) and link (rail) failures.

An average critical point of $\gamma_0 = 65\%$ means that it is difficult to propagate a cascade failure throughout a commuter rail network. Therefore, these rail networks are resilient against congestion that backs up traffic or faults that stop traffic altogether. On the other hand, an average link robustness value of 8% means that on average, removal of only 8% of the links (sections of track) can dismantle the network. It does not take much to separate the commuter system into disjoint islands. Because the structure of these systems is largely linear, removal of a single link can disable major components of the system. Figure 15.6 illustrates the linearity of a typical commuter rail. These hub-and-spoke networks typically contain one or two downtown hubs with outward spiraling arms connecting the suburbs to the city center.

Node robustness is considerably higher, but only because terminal nodes are considered blocking nodes. If terminal nodes are removed from the calculation of node robustness, the average is much lower than 31%. Nonetheless, removal of only a handful of stations (nodes) is likely to disable major components of these linear networks. Therefore, the major vulnerability of typical commuter rail systems is the lack of redundancy. Nearly every node and link is critical.

In most cases this weakness can be overcome by a small standby fleet of buses to substitute for sections of rail that

TABLE 15.1 Commuter/light rail systems are very fragile, even though they have low spectral radius, because they contain minimal node and link redundancy

Name	#Stations	#Links	Spectral radius ρ	Critical point γ_0 (%)	Node robustness (%)	Link robustness (%)	Max. paths	Avg. between (%)
SEPTA	80	90	3.34	45	32	12	1898	27
LA Metro	117	127	3.69	55	35	8	5424	26
Amtrak	55	66	3.06	61	60	18	1112	25
BART	45	46	2.51	113	13	4	1095	40
Boston MTA	125	125	3.13	50	15	0	4386	28
Average	84	91	3.15	65	31	8	2783	29

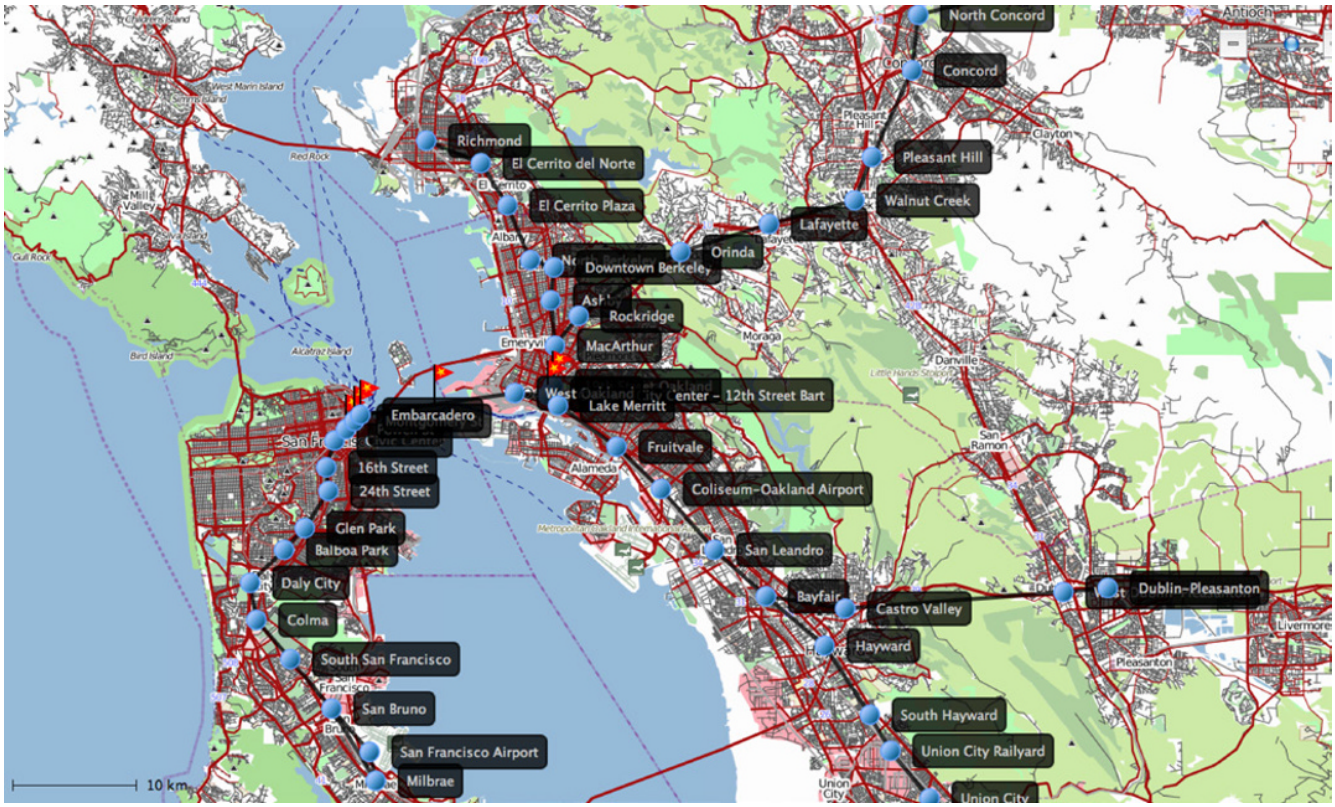


FIGURE 15.6 Bay Area Rapid Transit (BART) is a light rail commuter train serving over 2.5 million residents of the San Francisco Bay Area.

may be damaged accidentally or by terrorists. If a station or section of rail closes, buses can route around the closed asset. Similarly, if the trains or railcars are damaged, buses can be used temporarily to maintain business continuity. Long-term damage to a commuter rail system is unlikely and consequences are low.

15.4 AIR

Like highways and railroads, the commercial airline industry has been shaped by government regulation followed by deregulation. And, like the other modes of transportation, the government's regulate–deregulate legislation initially reduced competition and then increased competition. The tipping point came in 1978 when commercial air travel transitioned from a regulated sector resembling a public utility to a semi-deregulated industry resembling privately held infrastructure industries like energy, power, and communications. Prior to 1978, the Civil Aeronautics Board (CAB) determined which routes each airline flew and fixed airfares. After 1978 the industry became a market-driven business with government oversight.

A brief history of the regulatory phase of the cycle is as follows:

- The Kelly Act of 1925—private contractors allowed to carry the mail by air.
- The Air Commerce Act of 1926—promoted the development and stability of commercial aviation.
- The Civil Aeronautics Act of 1938—regulation of air transportation put under one federal agency.
- The Air Safety Board of 1938—created an independent body for the investigation of accidents.
- The Reorganization Act of 1940—split the Civil Aeronautics Authority into two agencies, the CAB and the Civil Aeronautics Administration (CAA).
- The Federal Airport Act of 1946—50/50% federal/local funding of airports.
- The Federal Aviation Act of 1958—transferred safety and air traffic control functions of the Civil Aeronautics Authority to the new Federal Aviation Agency (FAA).
- The DOT Act of 1966—created the DOT, renamed the Federal Aviation Agency as the Federal Aviation Administration, put the FAA under the DOT, and transferred the CAB's accident-investigation duties to the new and independent NTSB.
- Airport and Airways Development Act of 1970—established the Airport Development Aid Program (ADAP) and the Planning Grant Program to foster airport development.
- Airline Deregulation Act of 1978—created a competitive market and phased out the CAB's economic regulation of the airlines.

The *Airline Deregulation Act* of 1978 signed into law by President Jimmy Carter fell in line with the general trend of the late 1970s of pulling the United States out of recession by easing up on highly regulated sectors of the economy. Similar steps were being taken in other infrastructure sectors, such as the energy Public Utility Regulatory Policies Act (PURPA) of 1978. Economists of the time argued for competition as a means of economic stimulus, following the Yom Kippur war and oil embargo of 1973, which dramatically inflated fuel prices.

Technical advances also made it possible to remove price controls and encourage competition. For example, the wide-body passenger jet increased airline capacity and reduced passenger-mile costs. But prior to 1978, there were too many airliners flying too many routes, with too few passengers. The CAB responded to this crisis by initially increasing fares and reducing competition.

A 1975 report issued by the CAB concluded the airline industry was "naturally competitive, not monopolistic." Subsequently, the CAB began to loosen its grip on the industry and reversed course. The transition from regulation to deregulation began to sweep through different parts of the industry. In particular, air cargo and passenger segments began the transition from utility-like industry to a market-driven industry. Under the auspices of Congress, cargo carriers were allowed to fly any route they wanted and charge whatever the market would bear. In particular, overnight express delivery flourished. This allowed Federal Express—incorporated in 1971—to flourish and become the largest cargo airline company by weight.

The Airline Deregulation Act of 1978 applied the same free-market competition technique to passengers. By 1985 domestic route and rate restrictions were phased out by congressional mandate. Route and fare restrictions were eliminated, and the CAB was disbanded on January 1, 1985. Its remaining functions were transferred to the DOT. For example, the government is still responsible for granting landing rights in foreign countries to US carriers.

As a further example of deregulation, most airports are owned by private–public partnerships, but they can be owned and operated by private, public–private, local government, state government, or intergovernmental port authorities. Washington Dulles and National airports at one time were under federal ownership, but congressional legislation in 1986 transferred ownership to an airport authority. But military airports are still owned and operated by the federal government.

The Airline Deregulation Act ended government economic regulation of airline routes and rates, but not airline safety. Initially, the NTSB was part of the DOT. But in 1974, Congress reestablished the NTSB as a completely separate entity, outside the DOT, reasoning that "…No federal agency can properly perform such (investigatory) functions unless it is totally separate and independent from any other … agency of the United States."[5]

[5]http://www.ntsb.gov/about/history.html

Oddly, the accident rate and number of casualties increased for several decades after deregulation until roughly 1990. Subsequently, the accident rate and number of casualties has been declining (see Fig. 15.7a). While the probability distribution of casualties is not long-tailed, the average number of casualties per year is rather high at 1056 (see Fig. 15.7b). The PML risk profile obtained from the exceedence probability distribution reaches a maximum at 5.92 casualties. This says the maximum likely consequence of an airline casualty is 5.92. Figure 15.7b shows a rapid drop in PML risk beyond its peak.

The chance of dying in an air travel-related terrorist attack is extremely small. In fact, the database of recorded deaths due to a terrorist attack is too small to make statistical predictions—there have been only eight attacks due to suicides, sabotage, and terrorism since 1986 (see Table 15.2). There have been 582 deaths among 13.5 billion passengers and more than 800 million departures over the past 25–30 years. The odds of dying from a terrorist attack on an airplane or airport is too small to be meaningful. A deeper analysis is given at the end of this section.

15.4.1 Resilience of the Hub-and-Spoke Network

Regulation followed by deregulation resulted in today's hub-and-spoke network structure. This form of self-organization obviously increases travel efficiency and profits but sacrifices resiliency. The airline carriers claim that the hub-and-spoke system achieves higher percentage of filled seats and keeps passengers on the same carrier from end to end.

The hub-and-spoke structure reduces resiliency of the complex CIKR, because of high betweenness and hub centrality. Figure 15.8 shows the airport and route network formed by the most active 385 airports in the US domestic market. This network is extremely scale-free with a hub located at ATL. It consists of 934 domestic and foreign airports and 3888 routes.

The mean connectivity of this network is very low (8.33 connections), but the spectral radius is very high (50.8). ATL has 213 connections to other airports—both domestic and international. Link robustness is 76%, and node robustness is 98%. That is, 76% of the routes can be canceled and still have a connected system. This amounts to 2957 routes, leaving 931 critical links. But only 2% (64) airports hold the network together. Thus, there are 64 blocking nodes. These are critical to holding the network together.

This means cascade effects such as airport closures or route cancelations produce complex catastrophes when the probability of cascading from one airport to another exceeds 4.6%. Delays and cancelations are the consequences.

The top 10 blocking nodes (airports that keep the network from separating into disjoint islands) are also the most connected airports. In order, they are:

ATL
ORD
DFW

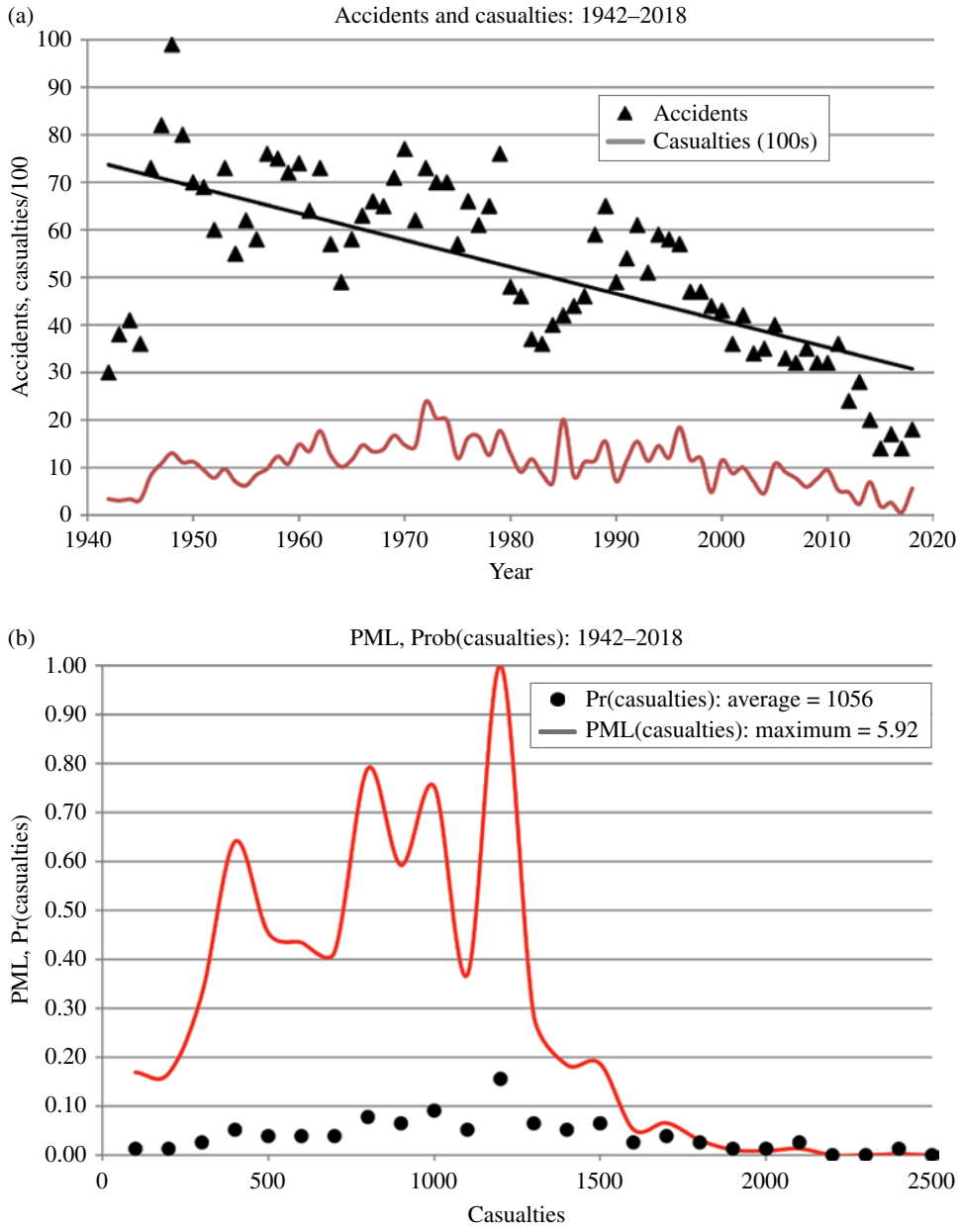


FIGURE 15.7 The casualty rate among all airlines carrying 14 or more people includes serious injuries and deaths. (a) Number of accidents and number of casualties in flights with 14 or more occupants began to fall in the 1990s globally. (b) Probability density of accidents obeys a normal distribution, approximately, rather than a long-tailed distribution. Average annual casualties totaled 1056, but maximum PML risk is relatively low at 5.92.

DEN
IAH
JFK
EWR
LAX
LAS

Betweenness analysis confirms the centrality and criticality of these airports. In fact, ATL is a super-spreader and super-connector because it is the most critical airport in terms of betweenness, connectivity, and number of high-betweenness routes. In order, the top ten betweenness nodes are:

ATL: Routes to ABY, AHN, AGS, ACY, AEX, ABE, ALB, ABR
ANC
DEN
HNL
JFK
ORD

DFW
SEA
HND
GRO

If betweenness and connectivity are combined, then the super-spreaders and super-connectors are:

ATL
DEN
ORD
DFW
JFK
EWR
LAX
IAH
ANC
SEA

TABLE 15.2 There have been 582 airliner deaths due to suicide, sabotage, and terrorism, out of 800,000,000 departures, since 1986

Date	Location	Deaths
April 2, 1986	Near Athens, Greece	4
December 7, 1987	San Luis Obispo, CA	43
December 21, 1988	Lockerbie, Scotland	270
April 7, 1994	Memphis, TN	0
September 11, 2001	New York, NY	92
September 11, 2001	New York, NY	65
September 11, 2001	Arlington, VA	64
September 11, 2001	Shanksville, PA	44
Total		582

Hardening these against cascade failure nearly eliminates complex CIKR collapse. In fact, the fractal dimension of the exceedence probability for collapse is very short-tailed, when vulnerability of all airports (except these) is 5%. Protecting these top 10 airports makes the US domestic network extremely resilient.

15.4.2 Security of Commercial Air Travel

In the United States, commercial airline security is the responsibility of the TSA. The TSA mission is simple, "to protect the nation's transportation systems and to ensure the freedom of movement for people and commerce." This mission applies to highways, railways, aviation, mass transit, pipelines, and postal/shipping systems. It is an ambitious goal requiring connections to domestic and international law enforcement and intelligence organizations. It is also a controversial program of security due to its intersection with privacy.

The TSA was moved from the US DOT to the DHS on March 9, 2003 in accordance with the Aviation and Transportation Security Act signed into law by President George W. Bush on November 19, 2001. Prior to its move to DHS, TSA was responsible for preventing bombings of airliners. On March 9, 1972, a dog named Brandy found a bomb on flight 12 min before it was to go off. As a consequence of Brandy's performance, President Nixon and the FAA created the Explosives Detection Canine Team Program, which still operates today.

TSA accounts for more than one-third of all 180,000 DHS employees. In addition to checking luggage and screening passengers at airports, TSA employees provide security-screening information to the commercial airliners through the Secure Flight information system. This

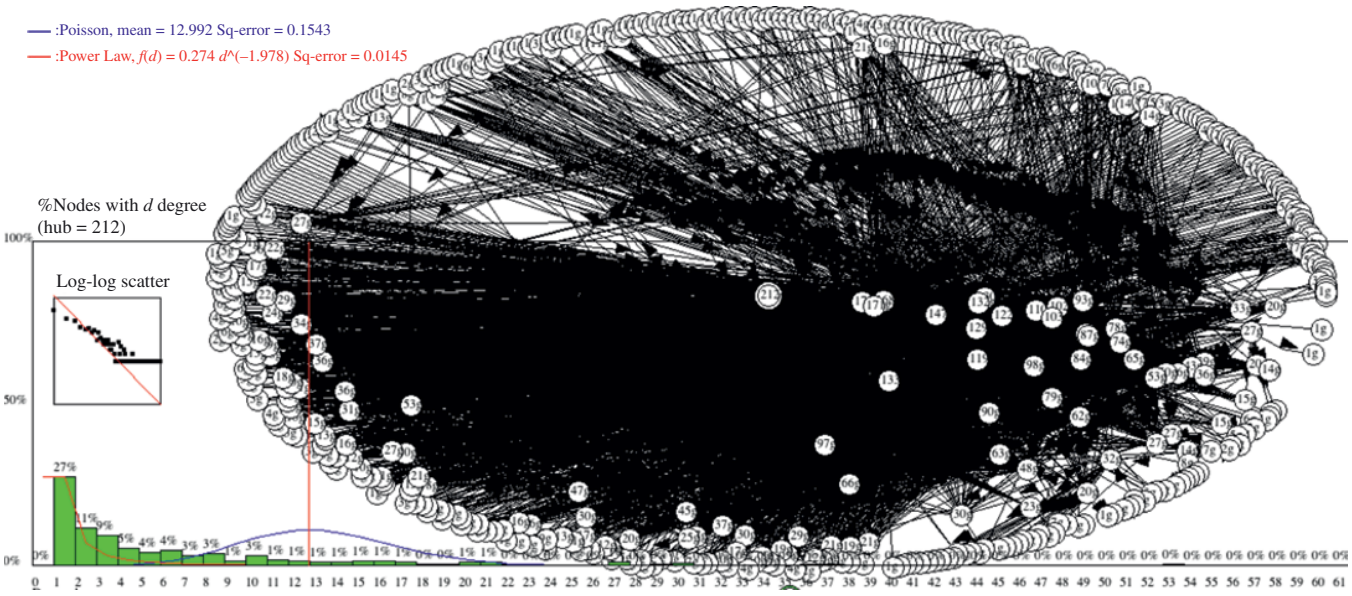


FIGURE 15.8 The primary airports and routes of the US domestic market form a scale-free network with spectral radius of 50.8.

controversial system links the *No Fly List* managed by TSA's TSC to each airliner's passenger boarding list. Airliners are alerted to matches between the No Fly List and suspected passengers. As of 2012, there were 21,000 names on the No Fly List.

The TSC collects passenger names, gender, date of birth, and frequent flyer information. The controversial Computer-Assisted Passenger Prescreening System (CAPPS) was scaled back after complaints that is violated the 1974 Privacy Act. This act was originally designed to protect medical and health information but has been used a number of times to prevent government collection and analysis of personal identification information—PII—which may be in violation of personal privacy. The act applies five core principles to government agencies to obey:

1. There should be no secret records held on citizens.
2. Citizens must be able to see the PII about them and how it is used.
3. The government must obtain written consent before PII collected for one purpose can be used for a different purpose.
4. Citizens must be allowed to correct their PII.
5. Government agencies are responsible for PII accuracy and must prevent its misuse.

CAPPS would have gone much further than Secure Flight. Although CAPPS II capabilities were not fully revealed, concerned citizens imagined a database that linked PII to records collected by other federal agencies to come up with total passenger information awareness. PII can be combined with the following records to determine much more private information:

CIA/FBI/NSA watch lists
Immigration/Customs/Border lists
Credit card/large banking transactions
Telephone/Internet contact with foreign suspects
Travel, car rental, hotel, restaurant histories
One-way tickets/no luggage

The General Accounting Office (GAO) July 22, 2005 report to Congress found TSA in violation of the Privacy Act of 1974:

> During the course of our ongoing review of the Secure Flight program, we found that TSA did not fully disclose to the public its use of personal information in its fall 2004 privacy notices as required by the Privacy Act. In particular, the public was not made fully aware of, nor had the opportunity to comment on, TSA's use of personal information drawn from commercial sources to test aspects of the Secure Flight. The Privacy Act provides safeguards against an invasion of privacy through the misuse of records by federal agencies and allows citizens to learn how their personal information is collected, maintained, used, and disseminated by the federal government. Specifically, a TSA contractor, acting on behalf of the agency, collected more than 100 million commercial data records containing personal information such as name, date of birth, and telephone number without informing the public. As a result of TSA's actions, the public did not receive the full protections of the Privacy Act.[6]

Nonetheless, TSA screeners employ a wide range of technologies to prevent terrorists from boarding airplanes. Electronic boarding passes not only expedite boarding but also allow TSA and the airlines to preview flyers ahead of time. A variety of biometrics, liquid scanners, explosive detection, and advanced imaging technologies are used to screen personal possessions and baggage. Image analysis performs face recognition and detects suspicious activities from cameras located throughout airports. These technologies may have prevented undisclosed attacks, but secrecy prevents confirming analysis.

15.4.3 How Safe and Secure Is Flying in the United States?

Commercial air travel risk assessment in the United States is complicated. A superficial analysis divides the number of injuries and fatalities by the number of boardings to obtain the likelihood of being injured or killed once you have boarded. An average of 82 passengers were injured or killed in an airline incident during the period 1993–2012, out of 824,000,000 domestic airline boardings in 2012.[7] Thus, a posteriori risk equals one injury or death per million passengers.

Another a posteriori estimate of risk of injury or death divides the number of injuries and deaths (1640) during 1993–2012 by the number of flights (201,662,785), to obtain the risk of flying on a commercial airliner, given that you board with numerous other passengers. This estimate is eight times higher but still very low. Odds of being injured or killed by flying are one in 123,000, according to this calculation.

Only eight incidents of air traffic fatalities due to sabotage, terrorist, or suicide attacks have been recorded during the period 1986–2012, including the 9/11 attacks. Total fatalities were 361. Therefore, the odds are one in 558,000 of dying from an intentional attack on a commercial airliner.

Compare airline travel risk with other forms of transportation. In 2016, 102 people died on US highways every day. Annually, 37,461 people died on US highways, 800 by

[6]http://www.gao.gov/new.items/d05864r.pdf
[7]http://www.ntsb.gov/data/aviation_stats.html

recreational boating, 759 by rail, 14 by pipeline accidents, and 444 by general aviation. There were 325 commercial airline fatalities in 2016 or 1 death per 10,769,230 boardings. By these measures, commercial airline travel is safe and secure.

15.5 AIRPORT GAMES

An incredible amount of effort and money has gone into securing the air transportation system since 2001. The response to the 9/11 terrorists attack verges on irrationality, given the statistical argument presented in Section 15.4.3. Nonetheless, airport security—and transportation security in general—illustrates one tactical approach to security that can be applied to other sectors. Lessons learned at airports may be applied elsewhere. GUARDS is one example, and Bayesian belief networks is another.

15.5.1 GUARDS

The GUARDS software program was developed for TSA by Milind Tambe and students at the University of Southern California to fend off terrorist attacks at airports [7]. TSA uses it "to allocate the TSA's limited resources across hundreds of security activities to provide protection at over 400 United States airports." It minimizes the risk of misappropriation of limited resources—people and equipment—in airports while randomizing defensive maneuvers by the defender. Theoretically, randomization makes it impossible for an attacker to predict where the defender will put resources to protect flights and passengers.

Figure 15.8 illustrates how GUARDS works. Consider the following two-person game. One player—the attacker—targets some asset such as a boarding gate, ticketing booth, or cargo area and attacks it with a weapon such as an IED or sarin gas. The attacker appears to the defender as a random agent in terms of time, place, and choice of weapon. This "attacker's connectivity of freedom" makes it very difficult to defend against with limited resources.

The defender has a limited number of resources in terms of people, machines, and countermeasures. These are shown as squares in Figure 15.9. Similarly, the defender has limited defensive measures, such as inspecting luggage and passengers. These are shown in Figure 15.9 as triangles. Finally, the defender has targets in common with the attacker, such as boarding areas and ticketing counters. These are shown as round objects in Figure 15.9.

The question posed by GUARDS is, "What strategy of resource allocation minimizes damage to the airport?" If the airport security defender allocates the two resources shown in Figure 15.9 to two targets employing two defensive measures, what happens if the attacker attacks a different target? GUARDS simulates the play by randomly assigning resources and defenses to targets and keeping score. When the defender successfully blocks an attack on a target by guessing correctly, GUARDS awards a point to the defender. When the attacker successfully attacks an unprotected target, GUARDS rewards the attacker.

The result is a mixed strategy. The combination of resources, defenses, and target selections that yield the highest score to the defender is used more often than the lower-scoring combinations. One strategy may be (randomly) applied 67% of the time and another applied 33%. Randomization makes up for the shortage of resources, but it still allows a small chance of success by a lucky attacker.

The risk of a successful attack diminishes as more resources are added to the defender's list of people, machines, and defenses. Therefore, the defender has a choice to make: spend more on security to lower risk or

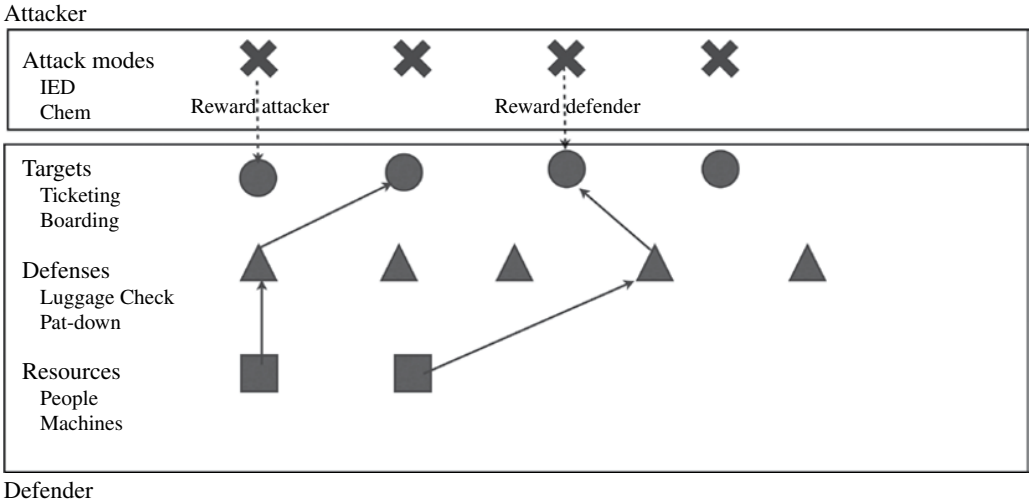


FIGURE 15.9 GUARDS is randomizing software that uses game theory to allocate limited defensive resources at airports.

spend less and suffer the consequences. GUARDS simply calculates the risk benefit—policy-makers must decide how much risk to allow.

The developers of GUARDS have also developed a similar game theoretic tool for the US Coast Guard called PROTECT. The idea is the same, simulate competition between two players—an attacker and a defender—to find optimal mixed strategies that minimize risk.

15.5.2 Bayesian Belief Networks

An alternative approach to risk minimization is based on Bayesian belief networks (BN) described in Chapter 2 and Appendix B. Bayesian networks are software tools for understanding *conditional risk*. As evidence accumulates, the likelihood of an attack either increases or decreases according to a network model of the attacker's intent and capability. A hypothetical airport scenario is used here to illustrate (see Fig. 15.10).

As described in Chapter 2, a BN represents the intentions and capabilities of actors in terms of propositions that are either true or false. Propositions are thought to be true with a certain connectivity of certainty, represented by a probability. As belief increases, so does the probability that the proposition holds. Belief increases or decreases according to evidence collected by observation or guesswork. An observation that supports a proposition raises the probability that the proposition is true.

Propositions interact with one another according to a network of influences (see Fig. 15.10), which contains a BN with five propositions:

Has luggage?
Has two-way ticket?
Has IED skill?
Has intent?
Dangerous?

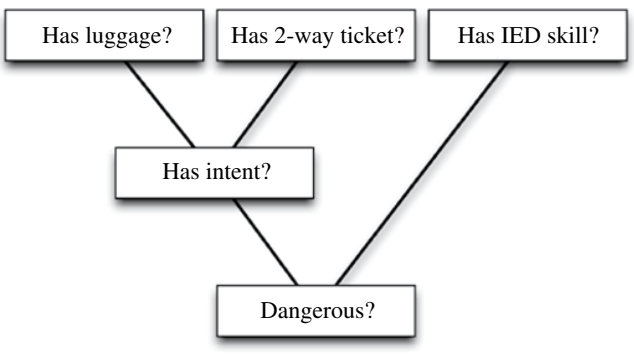


FIGURE 15.10 Bayesian belief network for airport security uses evidence to decide if a passenger poses a danger to passengers or flights.

The first three propositions are considered evidence. If a passenger has *no* luggage, intent is increased. If the passenger has *no* two-way ticket, intent is increased even more. For example, suppose one percent of all passengers have no luggage and 5% purchase one-way tickets. How much do we believe he or she intends to do harm? An estimate of this likelihood goes into "Has intent?" proposition and either increases or decreases the likelihood that the passenger has an evil intent.

Similarly, if the passenger has explosives expertise, he or she is likely to have IED-building capability. This belief is combined with the output from "Has intent?" proposition to calculate an estimate of the likelihood that the particular passenger is dangerous. Thus, the "Dangerous?" proposition depends on evidence provided by each input proposition.

Bayesian networks like the one in Figure 15.10 are only as good as the data entered into the software that computes Bayesian conditional probabilities. Someone must design a meaningful BN and populate it with probabilities, and someone else must populate the input propositions with evidence in real time. Both of these limit the usefulness of Bayesian networks.

15.6 EXERCISES

1. Which of the following is *not* part of the transportation sector?
   a. Ocean vessels
   b. Oil pipelines
   c. Gas pipelines
   d. Highways
   e. Commercial airlines
2. Intermodal transportation does *not* include:
   a. Seaports
   b. Coast Guard cutters
   c. Metropolitan buses
   d. Automobiles
   e. Trains
3. Which agency is *not* part of the Department of Transportation?
   a. FAA
   b. FRA
   c. FTA
   d. NTSB
   e. St. Lawrence Seaway Development Corporation
4. The Interstate Highway contributed 31% to the annual productivity growth of the United States in the 1950s. Subsequently, the Interstate Highway system:
   a. Increased productivity growth by 50%
   b. Increased productivity growth by 75%
   c. Decreased productivity growth by 7%
   d. Contributed 7% in the 1980s
   e. Contributed nothing to annual productivity

5. The resiliency of the National Highway System (NHS) is:
   a. Very high
   b. Very low
   c. Low, because bridges are critical nodes
   d. Similar to the resiliency of the power grid
   e. Similar to the resiliency of the Internet
6. The notion of big business and government regulation emerged from which infrastructure industry below?
   a. Commercial air travel
   b. The Internet
   c. The Railroads
   d. The National Highway System
   e. The gas and oil pipeline industry
7. Government regulation began with which of the following?
   a. Elkins Act of 1903
   b. Sherman Antitrust Act of 1890
   c. World War I
   d. Transportation Act of 1920
   e. Interstate Commerce Act of 1887
8. A tipping point in the regulation of passenger rail was reached when:
   a. The government nationalized the railroads in WWI.
   b. Penn Central Railroad failed.
   c. Amtrak was incorporated.
   d. Vanderbilt monopolized the early railroad industry.
   e. President Carter was elected President.
9. The history of infrastructure industry regulation has traced a cycle characterized by:
   a. Regulation followed by deregulation
   b. Mild regulation followed by more regulation
   c. Antitrust legislation aimed at breaking up monopolies
   d. Regulation of big businesses followed by all businesses
   e. Technology diffusion
10. The (intermodal) transportation hub of the United States is:
    a. Los Angeles and the LA–Long Beach port
    b. Atlanta and ATL
    c. New York
    d. Chicago
    e. Boston
11. The commercial air travel sector was regulated in 1926 and then deregulated commencing in:
    a. 1980
    b. 1981
    c. 1978
    d. 2001
    e. 2011
12. The odds of dying because of a terrorist attack on an airplane or airport is:
    a. Very high
    b. High, but lower than before 911
    c. Very low
    d. About the same as automobile accident rates
    e. Increasing since 9/11
13. The domestic airport and route network of the United States has a hub located at:
    a. Atlanta
    b. Chicago
    c. New York JFK
    d. New Jersey EWR
    e. Los Angeles LAX
14. The TSA's first line of defense against terrorist attacks is:
    a. CAPPS II
    b. Secure Flight
    c. Bayesian networks
    d. The 1974 Privacy Act
    e. X-ray machines
15. GUARDS is software for:
    a. Preventing terrorists from getting on airplanes
    b. Game theory software for allocating limited resources
    c. Bayesian network software for allocating limited resources
    d. A database of frequent flyers
    e. A No Fly List

15.7 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. Why do airline casualties obey a normal distribution instead of a long-tailed distribution?
B. Why is the spectral radius of the Interstate Highway System close to 4.0? The mean connectivity of the network in Figure 15.4 is 2.73. What does this mean in terms of resilience?
C. Why were there many more terrorist attacks on trains in the period of 1998–2003 than on commercial airliners during the period 1986–2012?
D. Figure 15.3 shows the cost of building and maintaining the Interstate Highway System rising from 0.05 to 0.30% of GDP over its first 50 years. Assuming the exponential rise continues, which theory of catastrophe does this illustrate? Use your answer to predict the future resilience of the network.
E. What was the first regulation imposed on American business by the US Congress, and why was it enacted? Compare your answer to the 1998 decision by the courts that Microsoft was a monopoly that should be broken into separate companies.

REFERENCES

[1] The Economic Impact of the Interstate Highway System. *Future Options for the National System of Interstate and Defense Highways*. NCHRP Project 20-24 (52), 2006.

[2] Gordon, P., Richardson, H. W., Moore, J. E., Pan, Q., Park, J., Cho, S., Cho, J., Jun, E., and Nguyen, C. *TransNIEMO: Economic Impact Analysis Using a Model of Consistent Interregional Economic and Highway Network Equilibria*, Los Angeles, CA: Final Report to the Center for Risk and Economic Analysis of Terrorism Events (CREATE), 2010.

[3] Thompson, L. S. *Regulatory Developments in the U.S.: History and Philosophy*, The World Bank, March 2000.

[4] Riley, J. *Terrorism and Rail Security*, Testimony presented to the Senate Commerce, Science, and Transportation Committee on March 23, 2004. RAND Corporation, Report No. CT-224, March 2004. Available at www.rand.org. Accessed July 3, 2014, pp. 3.

[5] Homeland Infrastructure Threat & Risk Analysis Center (HITRAC), *The Terrorist Threat to the U.S. Commercial Passenger and Freight Rail System*. May 24 2006. Available at http://abcnews.go.com/images/WNT/terrorist_threat_us_rail_system.pdf. Accessed July 3, 2014, pp. 3.

[6] Wilson, J. M., Jackson, B. A., Eisman, M., Steinberg, P., and Jack Riley, K. *Securing America's Passenger-Rail Systems*, Santa Monica: RAND Corporation, 2007.

[7] Pita, J., Tambe, M., Kiekintveld, C., Cullen, S., and Steigerwald, E. *GUARDS—Innovative Application of Game Theory for National Airport Security*. *International Joint Conference on Artificial Intelligence* (IJCAI), 2011, Barcelona, Spain.

16

SUPPLY CHAINS

A supply chain is defined here as a network of organizations, people, and assets used to move a product or service from supplier to customer. Nodes are generally factories, warehouses, and ports. Links are trucking, railway, and shipping routes. These are big, globe-circling networks that connect more than 178 countries to one another through a complex set of trade agreements, international laws, and customs agencies.

The wealth and security of every nation depends on robust import–export trade, and trade depends on an efficient and friction-free global supply chain. Ports are the infrastructure that makes the friction-free supply chain successful. Therefore, ports and shipping are critical nodes in this complex CIKR sector. Supply chain security, maritime domain security, and global intermodal transportation are nearly synonymous terms for the same thing. Regardless of the term used, supply chain security is of the highest concern among countries threatened by terrorists and criminals.

In the United States, the Department of Homeland Security (DHS) is responsible for establishing maritime security standards and agreements among participating countries. Operationally, security is enforced by a combination of TSA, Customs and Border Protection (CBP), and the US Coast Guard (USCG)—all components of the massive DHS. DHS has adopted a military-style layered strategy whereby agents are placed on-site in foreign ports, cargo is encapsulated in sealed containers, and inspections are conducted at several points along the chain. Layer one starts in another country, and subsequent layers continue with inspections at sea and destination ports.

Generally, supply chain management and maritime security has evolved from innovation in both shipping and security:

- *The tilted globe*: Globalization is driven by a tilt in wealth and labor. Over 62% of labor is located in the South and East, principally in Asia and Africa, while 62% of the wealth is located in the North and West, principally Europe and North America. This tilt drives globalization, as manufacturing seeks out cheap labor and wealth seeks out new markets abroad. The result is unprecedented demand for supply chain management and security. Thus, the tilted globe is driving rapid growth in ports and intermodal shipping.
- *Father of containerization*: Malcom McLean commercialized the modern form of intermodal cargo shipping called *containerized shipping* in the late 1950s. His innovative Sea-Land Services Corporation established the standard TEU (twenty-foot equivalent unit) shipping container, which transformed ports into automated exchange points in the global intermodal transportation network. Containerization radically reduced the cost of shipping from dollars to pennies and enabled subsequent globalization.
- *ISPS and CSI*: Security of the global supply chain is focused on ports. The International Ship and Port Facility Security (ISPS) code proposed and enforced by the United Nation's International Maritime Organization and the Container Security Initiative (CSI) that places US customs inspectors in foreign ports around the

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

world define supply chain security for the entire world. Adopted by 152 nations in 2004, these standards and practices form an encapsulated supply chain similar in concept to a trusted path in the information technology (IT) sector.

- *Security encapsulation*: Containers are sealed and inspected at several points along the supply chain, manually, and by X-ray and gamma ray machines. The Vehicle and Cargo Inspection System (VACIS) and Radiation Portal Monitor (RPM) scan containers as they leave a port and again when they enter a port.
- *Self-organized criticality*: The global supply chain is by definition self-organized to the point of reaching its critical point. All inefficient, redundant, and backup robustness has been eliminated in the name of cost reduction and speed. Therefore, it is an extremely fragile system. One failure such as the tsunami and nuclear power plant destruction at Fukushima, Japan, can severely damage the system. Consequences are extremely high in terms of lost product, lost productivity, and lost time.
- *Massive hubs*: Self-organization in the supply chain shows up as extremely large—and getting larger—ports. Because of the correlation between wealth and trade, emerging nations are competing with one another to build ever-larger ports. Shanghai recently overtook Singapore as the largest port in the world. Dubai is vying to replace Shanghai. The largest port in the United States is Los Angeles–Long Beach, followed closely by New York–New Jersey. Size drives self-organization, because big ports are more efficient than small ports.
- *Chokepoints*: Self-organization is also evident in the network formed by ports and routes. The most critical routes and ports are the most traveled and connected: Panama Canal, Suez Canal, Shanghai port, Singapore port, Antwerp port, and so on. The energy supply chain chokepoints are the Strait of Hormuz, Suez Canal, Bab el Mandeb, Turkish Straits, Danish Straits, and Panama Canal.
- *Trade equals wealth*: Gross domestic product (GDP) is highly correlated with import–export volume—the richest nations are also the largest traders. This is due to comparative advantage. Economic disruption can spread like a contagious disease through import–export flows among trading partners, introducing nonlinear fluctuations in the economy of distant nations. A common source of economic disruption is a Minsky moment—a type of paradox of enrichment that occurs when a segment of a country's economy exceeds its carrying capacity.
- *WTW*: The World Trade Web (WTW) is a network consisting of countries (nodes) and their import–export flows (links). Economic disruption in one country spreads to adjacent countries in the WTW according to this rule: a disruption in a large, highly connected trading nation can disrupt the economy of a smaller nation, but not the reverse. Economic disruption in a minimally connected nation has little impact on larger economies. Therefore, large trading nations are more secure than small trading nations, but small trading nations are susceptible to disruptions from large traders.
- *Port risk*: In the United States, the Coast Guard is responsible for port security. Maritime Security Risk Analysis Model (MSRAM) is a methodology and software tool used by the USCG to assess risk in ports. MSRAM extends the basic probabilistic risk assessment formula, R = TVC, in a number of novel ways—it is scenario based and refines the definition of threat, vulnerability, and consequence. The output from MSRAM is a risk index number for port assets considered critical. Resources are allocated on the basis of risk ranking.
- *Resource allocation*: The USCG uses a game-theoretic method and software tool called Port Resilience Operational/Tactical Enforcement to Combat Terrorism (PROTECT) to randomize and schedule Coast Guard patrols. PROTECT uses the risk index produced by MSRAM to put a value on targets that may be of interest to terrorists and criminals. PROTECT's implementation of Stackelberg competition calculates the best mixed strategy such that risk is minimized.
- *Trusted paths and secure supply chains*: In many respects, supply chain security is conceptually similar to trusted path security in the IT sector. Both sectors use a strategy of encapsulation and containment to harden their infrastructure. The unit of supply chain security is the sealed container. The unit of IT security is the encrypted packet. Both strategies transcend international boundaries and, in some cases, conflict with local regulations and laws. For example, supply chain security may violate local labor laws, and Internet privacy settings may violate local privacy laws. Therefore, both sectors are likely to undergo major changes over the next few years.

16.1 THE WORLD IS FLAT, BUT TILTED

Thomas Friedman wrote about globalization in *The World Is Flat*, but he missed an important point—the world is also *tilted* [1]. Over 62% of the world's working-age citizens live in the South and East—China, India, South America, and Africa. But over 62% of the world's wealth is concentrated in the North and West—North America and Western Europe. With only 5% of the labor, but 62% of the money, it is no

wonder that manufacturing is heading South and East, while products are moving North and West. Capital seeks out labor and profit margins, while consumption seeks out manufactured goods at low prices. Thus, *comparative advantage* drives globalization, which in turn depends on low-cost, efficient, and fast supply chains. A tilted world inevitably discovers globalization and invents intermodal transportation to provide goods and services throughout the world. Thus, supply chains emerged as a consequence of the tilt.

A tablet computer or cell phone ordered by a consumer in Germany, France, or the United States is manufactured in China and ends up in a consumer's hands only a few days later. As soon as the cell phone, say, leaves the factory, it is enclosed within a security layer much like an Internet packet is encapsulated in a *trusted path*. Only in the case of cargo the trusted path is a transportation network encapsulated in a series of security layers. The first layer is a sealed and locked shipping container. Subsequent layers are provided by the air or sea shipping company, with the oversight of the US government. The final layer is reached at the destination port, where the product is distributed to a store or consumer through another chain of transportation modes.

A Chinese trucking company arrives at the factory and loads the requested cell phone, for example, into a 20- or 40-foot container—called a TEU—along with thousands of other cell phones. The trucker bolts the container shut and stamps it with a security seal and transports it to the nearest departure port. The TEU will remain sealed until reaching a distribution warehouse in the United States or Europe. Thus, the first layer begins at the point of origin and continues through the trucking and transfer process as the product moves through the intermodal transportation system described in Chapter 15.

An intermediary shipping company determines the most economical and swift way to transport the container to the United States and fills out a manifest containing a list of contents, origination and destination point, and billing information. This manifest is submitted to the US government 24 h prior to shipping. A number of government agencies—USCG, TSA, and CBP—perform a risk assessment on the transaction prior to the container boarding an airplane or ship. They do this by looking at intelligence information, country of origin, shipper, and so on. Over 12 million containers enter the United States every year, so this is a major undertaking.

If the shipment is destined for the United States by air, it flies from China to Korea and then on to Anchorage, Alaska. From Alaska, it may go to Seattle, San Francisco, or Los Angeles if it is delivered to the western part of the United States, or to Louisville or Memphis, if delivered to the Eastern United States. If the container moves by ship, it is loaded onto a large container ship with upwards of 10,000 other TEUs and heads across the Pacific Ocean.

The transport layer of security is enforced by a rigorous set of standards proposed and enforced by the United Nation's International Maritime Organization. ISPS was adopted in 2004 by 152 nations. It is part of the DHS's CSI that posts US customs officials in foreign ports around the world. Over 20,000 ports and 55,000 ships fall under the CSI security umbrella. This layered strategy was adapted from the military strategy called a "layered defense."

96 hours before arrival at the Los Angeles–Long Beach port, the largest port in the United States, the captain of the ship must report to the USCG—who is on board the ship and what products are listed on the manifests for all of the containers on the ship. Once it arrives at LA–Long Beach, it enters a third security layer within the port. Port authorities as well as federal, state, and local law enforcement authorities manage this layer of security.

Containers are inspected before contents are distributed. A random sample of containers—or suspicious containers—may be opened and inspected in more detail. Manual inspection may take 6–40 h, depending on contents and intelligence information. A small sample is inspected by VACIS—a machine that scans contents of containers using gamma rays to look inside the container without opening it.

Finally, the container is loaded onto a train or truck and leaves the port. As it leaves, the container is once again scanned by RPM to detect radioactive material that may be inside. This is to prevent a dirty bomb from entering the country. At its destination, the container is opened and its contents are distributed to warehouses, stores, and consumers. Once in a consumer's hands, the product leaves the layered defense.

16.1.1 Supply-Side Supply

This incredibly secure and efficient intermodal transportation system is only the final step of a much larger supply chain network that encircles the globe. It is the demand tail that follows a much larger and more complex supply head. For example, the production of an Apple Inc. cell phone begins with an engineering design and ends with assembly of manufactured parts gathered from all around the world. Here is a (mostly) compete list of components needed to assemble an Apple iPhone 5C, circa 2013:

Apple design engineering	Apple custom chips
Apple A6 processor	Apple software
Qualcomm modem	Qualcomm transceiver
Elpida RAM	Toshiba flash memory
Cirrus audio codec	Qualcomm power management
Corning gorilla glass	iSight camera
Broadcom touch screen controller	TDK battery
Murata/Broadcom Wi-Fi module	Skyworks logic board
Avago A7900	TriQuint TQM6M6224
Hon Hai Precision manufacturing (Foxconn)	

These designs and parts from all over the world have to be transported efficiently and quickly from country of origin to China to be manufactured and then distributed back across the globe to consumers. Efficiency and speed is essential. Transportation once amounted to 10–12% of the cost of a consumer product, but now transportation amounts to only pennies of a product's cost. For example, the cost to transport an iPhone from China to the United States is less than $0.15. To ship a large-screen TV from China to Europe by boat, it cost as little as $4.00.

16.1.2 The Father of Containerization

Efficient and fast intermodal transportation of goods and services around the globe was not always possible before Malcom McLean commercialized containerized intermodal shipping. In 1956, the *Eisenhower Interstate Highway Act* and the first shipment of goods through McLean's containerized system revolutionized the US transportation system [2]. McLean is the father of container shipping—an invention he patented and freely licensed to all shipping companies to promote an efficient and fast method of transporting goods through the emerging intermodal transportation system of the post-WWII era.

McLean started McLean Trucking Company in North Carolina in 1937. He noticed the long delays and inefficient manual loading and unloading required each time cargo transferred from truck to ship and then back to truck again. This labor-intensive process introduced days of delay and drove the cost of cargo transportation ever upward. So in 1956 he tried a new approach—he left the loaded truck trailers intact and simply shipped them along with their cargo to the destination. As proof of concept, 58 trailers were directly rolled onto and off of a modified tanker ship named the *Ideal X* and delivered to Huston from the Newark port. The container era was born:

> But when Ideal X cast off from Berth 24 at the foot of Marsh Street in Port Newark, New Jersey, on April 26, 1956, and set a course for Houston, Texas, it was more than another tanker heading south in ballast to pick up additional product. Installed above the vessel's main deck was a special spar deck—a raised platform or porch—with longitudinal slots to which were attached the bodies of 58 trailer trucks. These were not trucks in any conventional sense—the 58 units had been detached from their running gear on the pier and had become containers. Arriving in Houston six days later, the 58 trailers were hoisted off Ideal X, attached to fresh running gear, and delivered to their intended destinations with no intermediate handling by longshoremen. [3]

McLean's shipping company—Sea-Land Services, Inc.—became legendary for transforming intermodal supply chains through container and port innovation. *Ideal X* transported 58 Lo–Lo (lift-on, lift-off) containers. The Sea-Land Services ship *Gateway City* accommodated 226 Lo–Lo containers. Today's ships and ports accommodate 10,000 Lo–Lo and Ro–Ro (roll-on, roll-off) containers in 20, 40, or 53 foot sizes. Bigger was not only better, but it was more profitable.

Sea-Land Services went through a series of owners—most recently as Maersk Shipping—and stimulated US–Asian trade beginning with the Pacific Triangle pioneered by McLean. During the Vietnam War, McLean's containerships transported military cargo from the United States to Vietnam. Rather than hauling empty containers back to the United States, McLean transported consumer goods from Japan on the backhaul. This evolved over decades into today's massive container ports along the Southeast Asian border. Today, Shanghai is the largest container port in the world handling over 32 million containers/year (see Fig. 16.1).

While containerization is an American invention, Figure 16.1a shows that Asia and the Middle East are the beneficiaries. Eight of the top 10 ports are located in Asia or the Middle East. Figure 16.1b compares the top 10 US ports with Shanghai—by far the largest port in the world in terms of TEU. Los Angeles–Long Beach is the largest port in the United States, followed by Houston–Galveston, New York–New Jersey, and New Orleans.

16.1.3 The Perils of Efficient Supply Chains

One missed shipment of parts can delay the construction of a Boeing airliner, resulting in millions of dollars of losses to the largest commercial airline manufacturer in the world. The Fukushima Daiichi event in 2011 damaged major parts of the global supply chain. The normal accident (earthquake, tsunami, nuclear power plant meltdown) damaged the electronics, automobile, and agricultural sectors of the Japanese economy, as well as manufacturers outside of Japan that depended on its products. The catastrophe in Japan caused supply chain managers to rethink "lean production" and "single sourcing."

The mantra of the twenty-first-century supply chain managers throughout the world is larger and faster containerships that drive costs downward. Just-in-time-inventory systems require efficient and inexpensive supply chains. Optimized efficiency is the number one security challenge, because efficiency is at odds with resilience and robustness. Efficiency removes redundancy and backup capacity. It sacrifices adaptability and flexibility for speed and low-cost handling. Efficiency means ever-increasing levels of *self-organized criticality*.

Optimized efficiency leading to self-organization is known as the *hourglass effect* in supply chain terminology [4]. Figure 16.2 illustrates the hourglass effect in a simple supply chain network. Note an hourglass shape is formed by placing the high-betweenness (and high-connectivity) nodes near the center and the input/output nodes at the edge of the

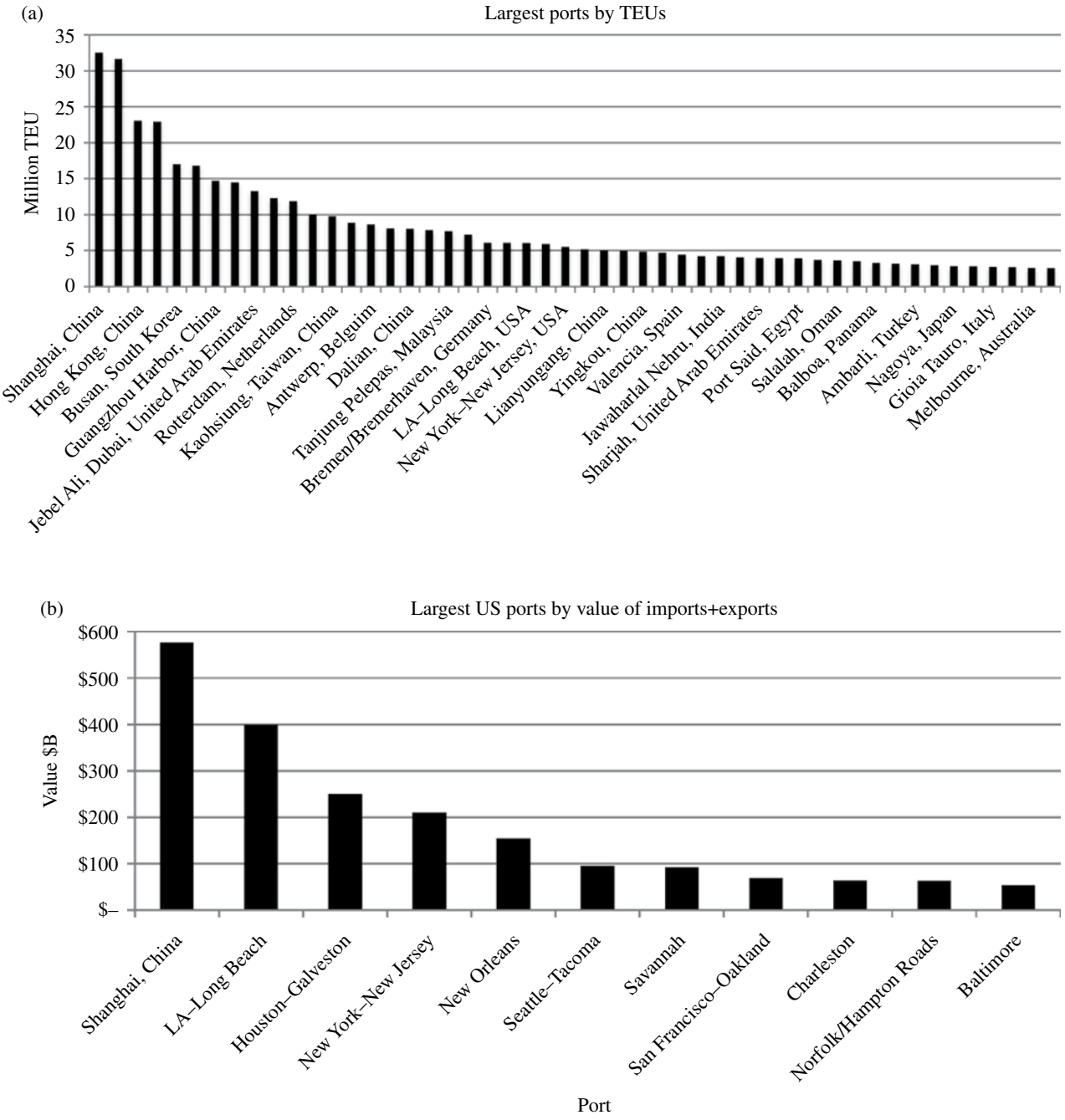


FIGURE 16.1 The largest ports in the world are mostly located in Asia. (a) Largest ports in the world by TEUs. (b) Largest ports in the United States by value of cargo, compared with Shanghai, circa 2013.

network. Commodity flow is funneled through these bottleneck nodes for reasons of efficiency.

Figure 16.2a ranks nodes c, d, and e highest in terms of betweenness centrality. Node d is also the hub, and node e is the second highest ranked node in terms of betweenness and connectivity. Figure 16.2b shows blocking nodes and links as also critical in terms of flow bottlenecks. There is a high correlation between blocking nodes and links and hourglass nodes. They are nearly identical in this example. Finally, the betweenness bottleneck analysis when one link is damaged shows there are alternate paths, but they become overloaded when one link is removed.

Self-organized criticality resulting in the hourglass effect increases betweenness indicating a potential bottleneck under stress. This is a by-product of optimization efficiency. Efficiency and speed have shaped the international supply

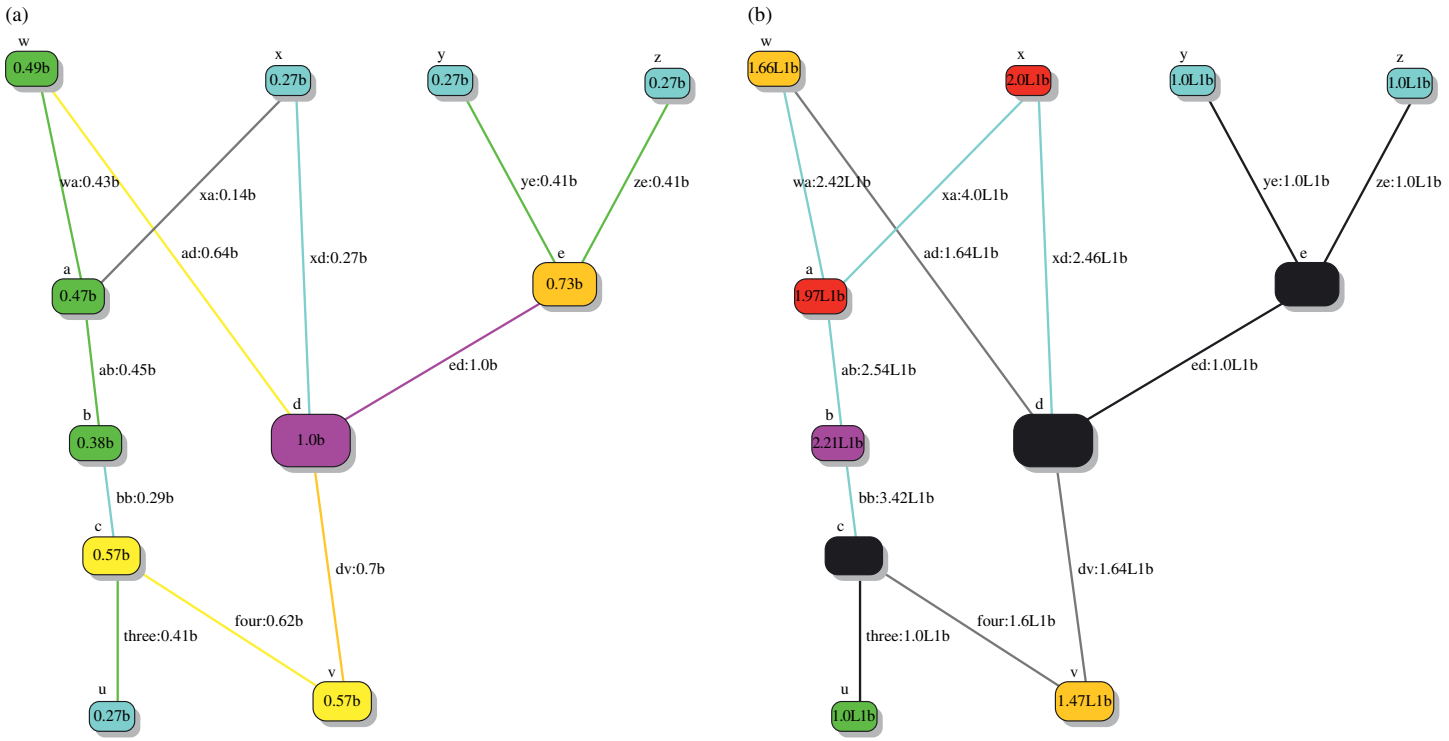


FIGURE 16.2 A typical optimized supply chain contains bottlenecks known as hourglass nodes. This is due to reduction of redundancy to save money. (a) Bottlenecks are revealed as high-betweenness nodes and links. (b) Blocking nodes and links are the ultimate bottlenecks. Hourglass nodes are typically highly connected high betweeners.

chain network, creating massive hubs like Shanghai and Los Angeles–Long Beach. But self-organization has also emerged in the form of high-betweenness nodes and links. For example, combining betweenness in the form of number of ships going in and out of the ports and connectivity in the form of number of connections to other ports, the 20 most critical ports and shipping routes are:

1. Panama Canal
2. Suez Canal
3. Shanghai
4. Singapore
5. Antwerp
6. Piraeus
7. Terneuzen
8. Plaquemines
9. Houston
10. Ijmuiden
11. Santos
12. Tianjin
13. New York and New Jersey
14. Europoort
15. Hamburg
16. Le Havre
17. St. Petersburg
18. Bremerhaven
19. Las Palmas
20. Barcelona

In addition, the *energy supply chain* is particularly critical to the security of the United States. Oil tankers and supply routes are highly vulnerable to disruption. High-betweenness chokepoints have evolved over the decades as oil tankers have increased efficiency by getting bigger and the most cost-effective shipping lanes have become routine. The most critical links and nodes in this supply chain along with the fraction of global oil supply passing through these chokepoints are:

Strait of Hormuz	20%
Suez Canal	4%
Bab el Mandeb	4%
Turkish Straits	3%
Danish Straits	3%
Panama Canal	1%

The story is similar for the shipment of food and other essential commodities. For example, the global share of grain exports of the top eight routes is:

Malacca	18%
Panama Canal	14%
Turkish Straits	12%
Gibraltar	10%
Suez	9%
Bab el Mandeb	8%
Dover	4%
Hormuz	4%

Source: Chatham House Maritime Analysis Tool; Chatham House (2017), resourcetrade.earth (2015 data).

16.2 THE WORLD TRADE WEB

Global tilt, global trade, and global supply chain networks are fragile systems. Like most other complex CIKR systems, they respond nonlinearly to shocks. Figure 16.3 shows the relationship between exports and GDP of the three largest traders in the world. The United States is by far the largest trader, followed by the European Union and China. These three regions of the world are also the largest economies on the planet. In all cases, there is a linear relationship between exports and GDP. That is, the more trading a country does, the wealthier it becomes. This is due to comparative advantage—the economics of trade.

In addition to a strong correlation between export volume and wealth, the graphs of Figure 16.3 contain nonlinearities brought on by economic disruption. For example, the largest nonlinearities in Figure 16.3 occurred in 2008–2010 because of the financial meltdown of 2008. Thus, the relationship between GDP and trade contains both linear and nonlinear components. When the global economy is stable, GDP grows with exports. When one or more of the traders encounter an economic disruption, the shock spreads like an epidemic to its trading partners. The economic contagion introduces a nonlinear response as shown in Figure 16.3.

The graphs of Figure 16.3 also plot the predicted impact of a disruption on GDP using a *paradox of enrichment* model. The result of trade and enrichment is shown as a dotted line and generally fits the actual GDP versus export curve with a root-mean-square error of 85% or more [5]. In economic theory, a *Minsky moment* can be precipitated by enrichment. These moments cause nonlinear fluctuations in GDP. In simple terms, some sector of the economy expands too fast and exceeds the carrying capacity of the economy. Too many people borrowing money to buy houses in the United States precipitated the 2008 financial meltdown, because homeownership exceeded the carrying capacity (roughly 65%) of the US economy in 2008. Similarly, "easy money" in other countries has enriched the economy beyond its carrying capacity.

The question posed here is, "What is the effect of an economic disruption in one country on another country?" That is, can disruptions spread from one country to another like a disease? If so, then the impact of an economic crisis in one country should have a measurable effect on trading

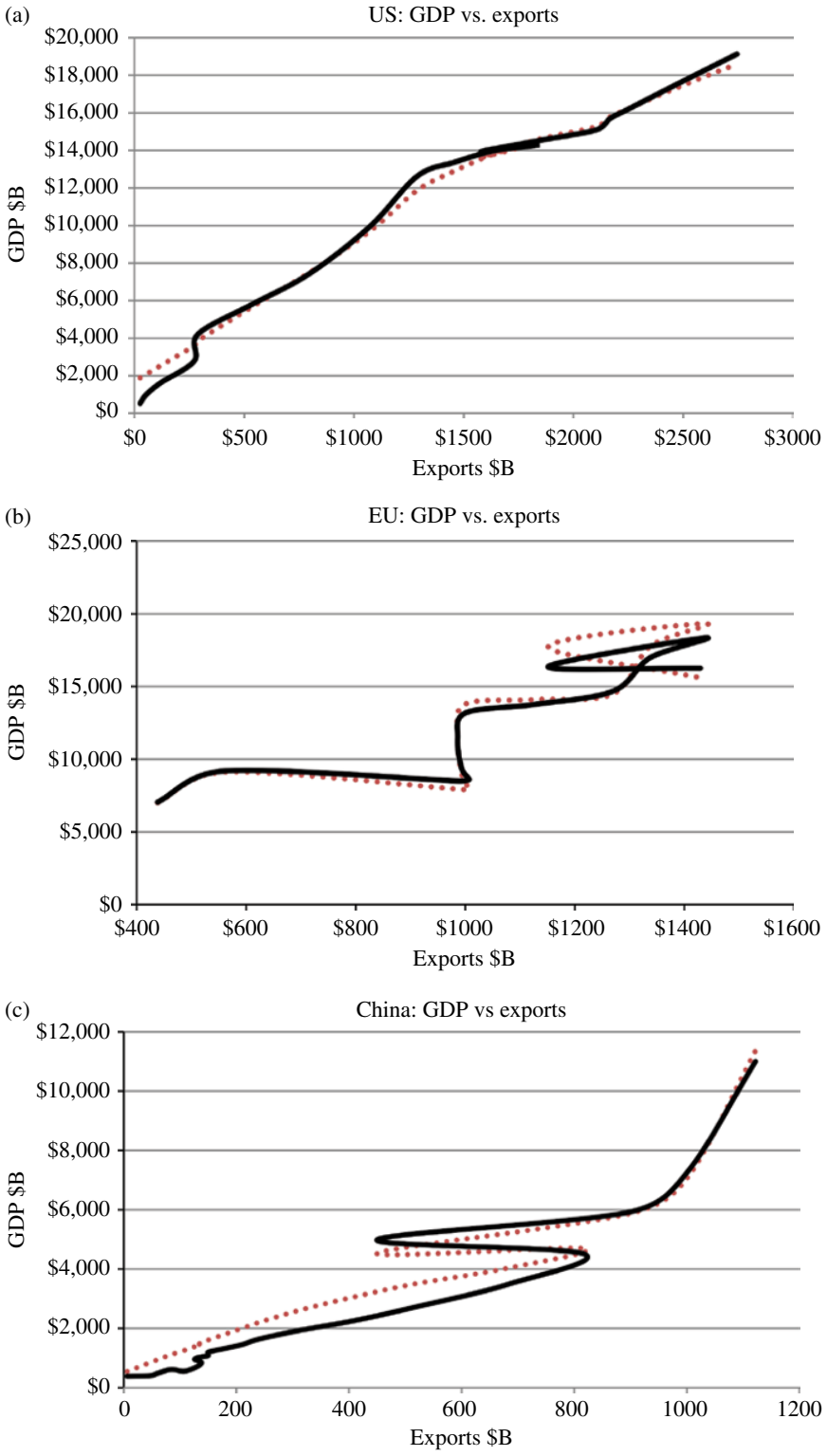


FIGURE 16.3 The GDP of the United States, the European Union (EU), and China is a nonlinear increasing function of exports: higher exports equal higher GDP. (a) US GDP versus exports 1960–2020. (b) EU GDP versus exports 1990–2010. (c) China GDP versus exports 1990–2012.

partners, even when the trading partner has a sound economy. If not, then a disruption in one country should have little impact on its trading partners. Do economies behave like contagious germs?

16.2.1 Economic Contagions

To address this question, researchers have constructed a WTW as shown in Figure 16.4. The nodes of this network are countries, with a value equal to their GDP. Links are trade relationships—imports and exports—given a value equal to the value of trade. For example, the United States was connected to 112 trading partners and had a GDP of $11,400 billion when Figure 16.4 was created.

The WTW analyzed here contains 178 countries, but only 93 have significant import–export links. Its mean connectivity is 6.37 links and its spectral radius is 15.96—modestly self-organized into a scale-free network. Nonetheless, the United States is at the center of this network with 112 connections. Average betweenness is low at 6.4% of 6444 paths through the US hub. The United States is the largest trading nation and is by far the most influential node in this network.

Table 16.1 lists the top 10 countries in each measure of structure. Clearly, the United States is central, with Singapore and other Asian countries close behind. Interestingly, Argentina ranks high in terms of betweenness, because it is an intermediary between South American countries and the rest of the world. The Netherlands ranks high in terms of connectivity because it has historically been a trading nation. The United States, however, is the "glue" that holds most of the network together.

A disruption in one node (country) should spread to its trading partners through lower trade. The disrupted economy should respond by decreasing its imports, which means that its trading partners sell fewer exports. A decline in exports should lower GDP, which sets into motion a normal accident—a chain of declining exports resulting in declining GDPs. But this does not always happen, as history records.

The *Tequila crisis*—a shocking devaluation of the Mexican peso in December 1994—produced a lot of bad press but failed to shock the rest of the world. It was caused by enrichment—a decade of hyperinflation, heavy government debt, ultralow oil prices, and a shortage of cash. It took Argentina down with it, but the rest of the world was virtually unscathed.

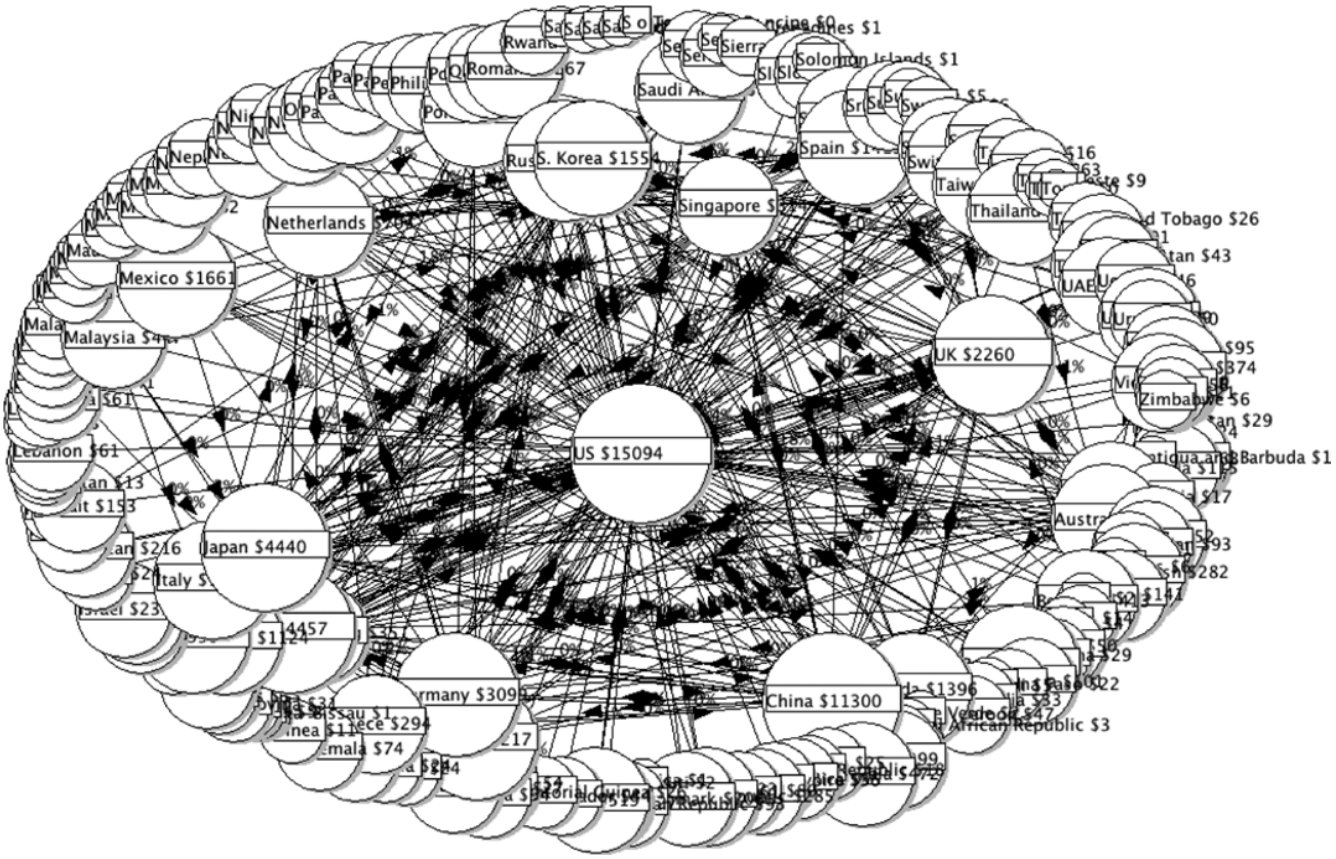


FIGURE 16.4 The World Trade Web is a network of 178 nations (nodes) and their 296 import–export trading relationships (links).

TABLE 16.1 Top 10 countries in the WTW ranked by network properties generally indicate that a handful of nations are central to world trade

Connectivity	Betweenness	Connectivity + betweenness	Link betweenness
United States	United States	United States	United States → Afghanistan
Singapore	Singapore	Singapore	United States → Algeria
United Kingdom	Russia	Russia	United States → Angola
Germany	China	China	United States → Azerbaijan
Russia	Honk Kong	United Kingdom	United States → Bolivia
China	Argentina	Hong Kong	United States → Chad
Japan	United Kingdom	Germany	United States → Chile
Hong Kong	India	Japan	United States → Austria
South Korea	South Korea	South Korea	United States → Columbia
Netherlands	Japan	Netherlands	United States → Congo

The 1997 *Asian flu* gripped much of Asia following a financial crisis in Thailand. High foreign debt crushed the Thai baht and bankrupted the government. The "Asian economic miracle" of the previous decade was fueled by financial enrichment. The Southeast Asian banks charged high interest rates, which attracted large investments, which in turn supercharged their economies. During the ramp-up, countries like Thailand, Malaysia, Indochina, Singapore, and South Korea grew at rates of 8–12% annually. But the Asian flu soon burned out before contaminating the entire global financial system. The Asian flu made a bigger impact than the Tequila crisis, but still failed to disrupt the US economy.

The Russian virus of 1998 also flared up and died out without capsizing the world's financial system. In fact, all of the shocks described here had disastrous effects on certain countries, but not others. Argentina, Venezuela, Thailand, and Mexico were heavily impacted, but US, Chinese, Indian, German, and UK economies were not. What determines the impact of financial shocks in one part of the world on the rest of the world?

Link robustness is high with 280(75%) of the trade links considered redundant. But node robustness is a different story. Only five blocking nodes hold the network together. Removal of any one of these countries separates the WTW into disjoint components—the United States, Hong Kong, Singapore, Russia, and Argentina. These are the "backbone traders" that the other countries depend on to keep the network in one piece. Together, these 5 countries make up 75% of the 296 trade links.

The fundamental resilience line of WTW points to a fragile network due to the small number of blocking nodes (5), large number of blocking links (34), and spread of economic loss due to economic failure of the United States and the other four blocking nodes. For example, an economic disruption starting with the United States spreads to 112 trading partners. A disruption starting with Argentina cascades through much of South America. Additionally, a handful of nations are so strongly connected that a disruption in one quickly spreads to an adjacent nation. Portugal and Spain, Czech Republic and Poland, Paraguay and Argentina, and Belgium and France are examples.

Resilience is maximized when the five blocking nodes are protected. The least resilient scenario occurs when the United States fails. An economic disruption in Argentina has a far less severe impact than a US disruption, but is slightly more damaging than a random disruption. We conclude that large traders like the United States have a large impact on the world, but smaller traders like Argentina have a smaller impact.

Other researchers have found similar results. Angeles Serrano and Marian Boguña of the Universitat de Barcelona, Barcelona, Spain, also showed that the WTW is a complex network [6]: it is *scale-free*, wired together like a *small world* (small diameter), and contains clusters of nodes representing regional trading partners. Fortunately, it is also rather resilient against shocks in GDP, because large stable economies with many trading partners tend to smooth out disruptions that emanate from smaller traders. In other words, the Barcelona investigators reasoned that large traders could disrupt small traders, but small traders could not disrupt large traders. Furthermore, big economies (and traders) spread contagions, while small economies (and traders) do not. Essentially, "big ships capsize small boats, but the reverse is not true."

The so-called rich club nations conduct more trade than the so-called emerging nations. Trade intensity is equal to the number of trade links connecting a country to other countries. Thus, we can rank countries according to their connectivity or network *connectivity*, because more links equals more trade. High-connectivity countries are big traders. Table 16.1 goes one step further—countries with both high connectivity and high betweenness have a heightened impact on lower-ranking trading partners.

Stefano Schiavo and colleagues at the University of Trento, Trento, Italy, claim, "International connectedness [alone] is *not* a relevant predictor of crisis intensity." In fact, they

concluded the reverse, "adverse shocks dissipate quicker" for countries with more trading partners. Rather than spreading financial contagion faster, a WTW country with many trading links tends to dissipate financial contagion.[1] In other words, connected and high-betweenness countries stabilize shocks created by lower-ranking countries.

Applying this newfound insight, the most and least vulnerable trading nations in the WTW are determined by number of trade links, for example, node connectivity. The most vulnerable countries in Figure 16.4 are Egypt, Iran, Taiwan, South Africa, Saudi Arabia, Thailand, Mexico, Poland, Brazil, and Spain. The least vulnerable are the United States, China, India, Japan, Germany, Russia, the United Kingdom, France, and Italy.

The connection between trade robustness and security places even more importance on ports, because ports are the economic hubs of international trade. Secure ports mean secure trade and secure trade means resilience. Trade robustness is a national security issue, because trading makes a nation more resilient against economic disruptions.

Therefore, port security becomes equivalent with supply chain security.

16.3 RISK ASSESSMENT

The gates of the largest planned settlement on Earth opened for business in February 2007 in the United Arab Emirates (UAE). Dubai World Central developed the 87,500-bed Logistics City expressly to run the Port of Jebel Ali. Jebel Ali is the world's largest man-made harbor and the largest port in the Middle East. To attract international businesses, UAE exempts tenants from taxes for 50 years and places no limits on the amount of money that can be moved in and out of the country. 800,000 Emirati citizens benefit from the work of 4.5 million immigrant workers. Dubai intends to become a supply chain giant, because economic power and security in the twenty-first century depends on supply chain superiority.

Economic success is tied so tightly to supply chain expertise that ports have become the most critical of critical infrastructures. They are the first layer of the layered strategy described earlier. This layer is implemented by the Transportation Worker Identity Credential (TWIC)—a security program instigated by the US DHS in 2004. It is managed jointly by the TSA and USCG and operated by Lockheed Martin Corporation. It affects 1.5 million workers. TWIC and CSI operate across borders and nations and envelope massive port authorities like Jebel Ali, Hong Kong, Shanghai, Los Angeles–Long Beach, New York–New Jersey, and Singapore.

Port security begins with prevention and risk reduction. TWIC and CSI are designed to reduce vulnerability and consequences, that is, risk. But how are these risks measured? What does a risk-informed decision-making strategy for supply chain management look like, and who is responsible for implementing the risk-informed strategy?

16.3.1 MSRAM

Port security begins with port risk assessment. All major ports under US control must be evaluated using the risk-informed decision-making tool called MSRAM, developed by a USCG team led by LCDR Brady Downs [7]. MSRAM is scenario based, meaning input data is elicited from hypothetical attacks called *modes*. As of 2010, 28,000 assets and 85,000 attack modes had been evaluated using MSRAM. This places MSRAM at the forefront of risk assessment tools available to assess critical infrastructure.

MSRAM implements the R = TVC probabilistic risk analysis method using scenarios (attack modes) to obtain estimates of T, V, and C. *Attack modes*—hypothetical terrorist or natural hazards—are classified as aerial, cyber, insider, landside, waterside, or combinations. Modes are matched with targets, which narrows down the range of values acceptable for T, V, and C.

MSRAM is unique in the way it validates risk. Port captains are responsible for conducting MSRAM assessments, but results are forwarded to USCG districts where they are compared with similar ports. Then the district results are forwarded to USCG headquarters where similar ports and assets are once again compared. In this way, risk values in one port are in line with values at other ports. MSRAM risk for port A is relative to risk at all other ports. This methodology is effective against "gaming the system" to receive more resources than competing ports.

Generally, threat T is an estimate of a specific threat against a specific target as determined by USCG intelligence. Threat–asset pairs are threat–target pairs in MSRAM. They are matched with estimates of consequence. Therefore, MSRAM risk is more accurately defined as a threat–asset–consequence triad. Additionally, vulnerability can be reduced by prevention, and consequence can be reduced by improvements in response. Therefore, a more lucid model of risk is given by the product of TVC and α, where α is a response mitigation factor:

$$R = TVC\alpha$$

T : threat

V : vulnerability

C : consequence

α : response

[1] "Higher interconnectedness reduces the severity of the crisis, as it allows adverse shocks to dissipate quicker. However, the systemic risk hypothesis cannot be completely dismissed and being central in the network, if the node is not a member of a rich club, puts the country in an adverse and risky position in times of crises. Finally, we find strong evidence of nonlinear effects."

MSRAM implements a number of innovations in applying this equation. For example, T is further decomposed into intent, capability, and geographic location. Intent is an estimate of the fervor of an attacker to do harm. For example, intent is exacerbated by hatred or current events reported in the news. Capability is a measure of expertise or access to resources and geographic location factors in the attractiveness of certain areas of the country. For example, someone skilled in building bombs has more capability than an unskilled terrorist, and large ports are more attractive to terrorists than small ports.

Vulnerability V is also decomposed into parts by MSRAM. Vulnerability increases with achievability and target fragility, but decreases with hardened owner/operator security practices. Effort by local law enforcement, owner/operator, and USCG tends to reduce V. Thus, V is roughly equivalent to Achievability + Target fragility – Security effort.

Similarly, consequence C is composed of three major components: primary, secondary, and response. Primary consequences are the result of deaths and injuries, direct economic impact, environmental impact, national security impact, and symbolic impact on the public. Secondary impact is generally due to economic losses due to fear of shopping, traveling, or going to work. Response is quantified by the α factor in the modified MSRAM equation above. Rapid and capable owner/operator, first responder, and USCG responses reduce C by an amount equal to α, where α is a number equal to or less than 1.

The USCG lists the following target types and attack modes: barge, building, infrastructure, key asset, and vessel. Attack modes are truck bomb, boar bomb, assault team, hijack team, swimmer/diver, malicious passenger, sabotage, and multiple bomb attack. Cyber, chemical, biological, and radiological attacks are listed as optional modes, as are aquatic and land mines.

Typical scenarios are a boat bomb attack on a ferry terminal in harbor, a car bomb on a facility in Houston, an attack on a cruise ship, and a bombing of a critical pipeline.

The risk index R of each scenario is compared to determine the rank order of threat–asset–consequence triads. Resources are allocated according to the risk index rank. Figure 16.5 shows how the USCG evaluates risk ranking. Every threat–asset–consequence triad is plotted on a scatter diagram divided into four quadrants. The upper left-hand quadrant represents highly likely, but low-consequence events. The lower right-hand quadrant represents unlikely, but high-consequence events. The four quadrants represent the four possible combinations of high and low likelihood of success versus consequence.

Risk indexes typically cluster along contours as shown in Figure 16.5. Risk group #1 contains highly likely high-consequence scenarios. Risk group #2 falls in the middle, and group #3 are unlikely low-consequence scenarios. Thus, assets are ranked according to their location on this scatter diagram.

As described in previous chapters, rank ordering risk is not guaranteed to minimize risk across multiple assets.

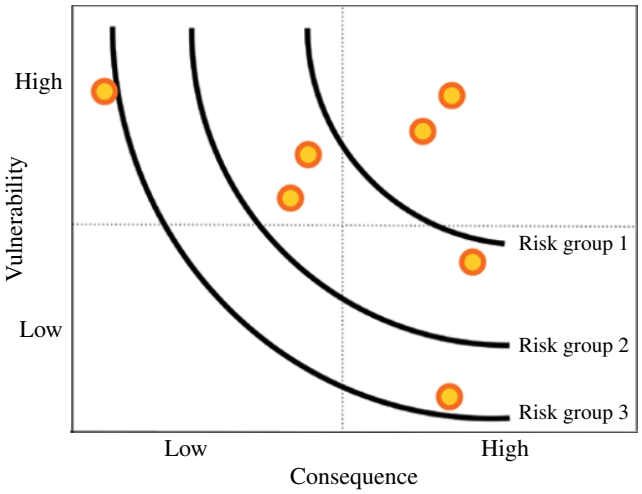


FIGURE 16.5 Risk ranking in MSRAM considers vulnerability and consequence.

However, allocation by ranking will reduce the maximum risk across a portfolio of threat–asset–consequence triads. This strategy is in line with the USCG policy.

16.3.2 PROTECT

PROTECT is a two-person Stackelberg game similar of TSA's GUARDS, described in Chapter 15. Its objective is to schedule randomized patrols for USCG cutters within harbors, waterways, and coastal shipping. Like GUARDS, PROTECT combines probabilistic risk with deterministic modeling to obtain an optimal mixed-strategy solution to allocating limited resources. The output from PROTECT is a weighted randomized list of patrols for Coast Guard cutters and personnel.

Figure 16.6 illustrates a simple PROTECT scenario for the Los Angeles–Long Beach port. Nodes represent targets and links represent patrol routes. The USCG has limited crew and boats to send to each target, while an attacker has unknown resources to apply to one or more targets. The challenge then is to schedule crew and boats to patrol this vast port in an optimal manner. PROTECT uses MSRAM risk index numbers to value each target and produces a randomized mixed-strategy patrol schedule. If the USCG successfully defends a target with risk index equal to 1000, it "wins" 1000 points. If the attacker successfully attacks the same target, it "wins" 1000 points. The objective of the game is to win as many points as possible.

For example, there are three routes shown in Figure 16.6. Assuming the USCG has only one boat and crew to deploy, the optimal mixed strategy below produces the most "wins" for the USCG:

Route #1 (Alameda): 50% of the time
Route #2 (Navy Way): 33% of the time
Route #3 (Desmond Bridge): 17% of the time

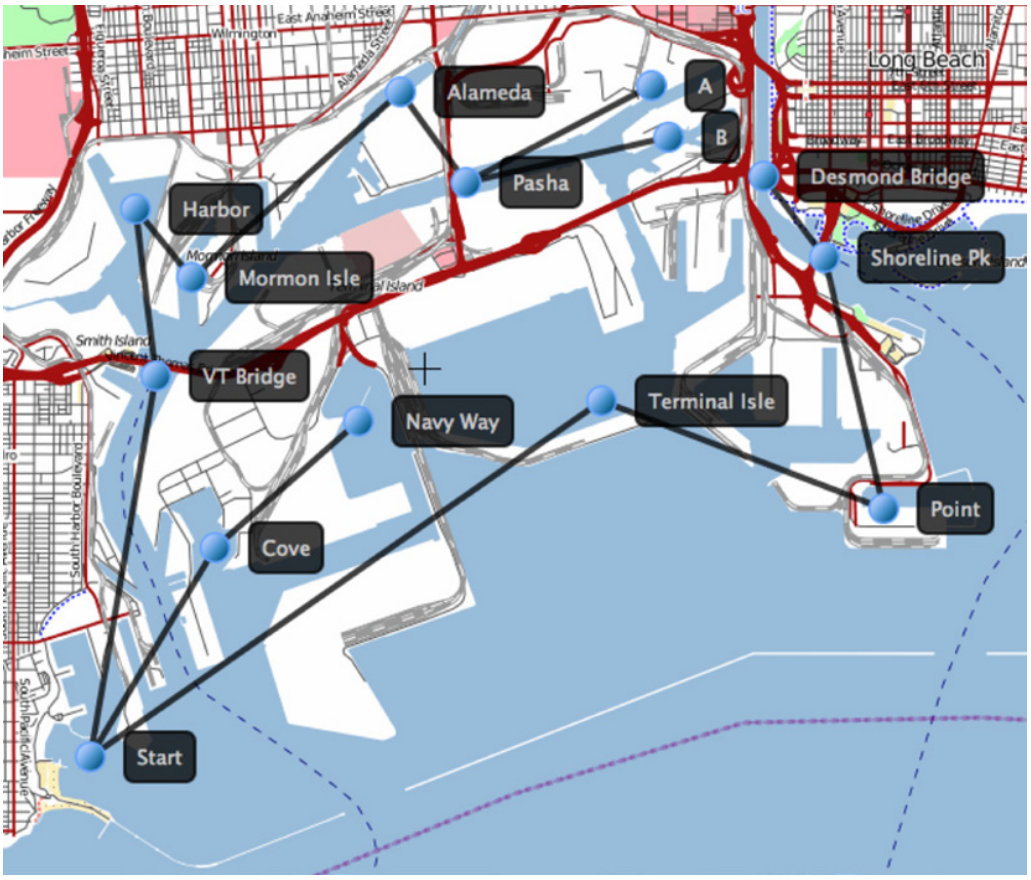


FIGURE 16.6 PROTECT uses Stackelberg competition to allocate limited resources to randomized patrols within ports and harbors.

Additionally, these routes are randomized so that an attacker cannot anticipate when a patrol might appear. Computerized randomization selects Route #1 50% of the time, Route #2 33% of the time, and Route #3 the remainder of the time. For example, if 10 patrols are possible in a 24 h period of time, 5 patrols follow Route #1, 3–4 follow Route #2, and 2–3 follow Route #3.

16.4 ANALYSIS

There is a striking similarity between the containment strategy of the global supply chain network and the Internet. As briefly mentioned above, the US strategy of enclosing ports and shipping routes mirrors the same approach of enclosing IT systems in trusted paths. In fact, computer input and output ports are the access points for hackers attempting to exploit the Internet, just as seaports are the access points for terrorists and criminals attempting to exploit the global supply chain.

In both sectors the strategy of containment and blocking transcends national borders. The US government-enforced ISPS code requires workers to undergo invasive security screenings that include criminal background checks, immigration checks, and intelligence-driven terrorism background checks. The TWIC process may violate local privacy and right-to-work laws, for example. In short, the ISPS enforces a kind of international law on top of local laws and regulations. Urban planner Deborah Cowen writes,

> Programs like the TWIC govern ports as exceptional seam spaces of economic flow that are literally outside the space of normal national law. [8]

Similarly, privacy and security regulations (and cultural memes) applied to the Internet may conflict with local regulations and laws. For example, privacy standards in the European Union are different than in the United States. Therefore, strategies for securing the Internet differ. How can these differences be resolved? In the case of ports and supply chains, the United States has exerted its power as the largest trading nation on the planet to assert strict control of maritime security. This level of assertiveness may not be permanent.

The case of Dubai Ports World (DP World)—a UAE state-run company that attempted to take over the operation of 22 US ports in 2007—illustrates the conflict between secure trusted path encapsulation and politics. Potential

ownership of the trusted path by an Arab state provoked congressional opposition on the basis of national security. Ironically, ownership by DP World would have improved security, because it would have encapsulated port security within the DP World network. Dubai is touted as a model for US port security, but the uninformed Congress and outcry from a biased public halted the deal, and DP World sold its interests.

Currently, these sociopolitical and economic factors are in the background. They have not surfaced as a headline, but at some point in the future, they may become important because of a type of political self-organized criticality that comes with liberal democracies. Political opposition to government often builds through a sociopolitical process similar to preferential attachment. The TSA has already experienced pressure for how it conducts inspections in airports. Might the TSA—and USCG—come under fire in the future because of its transnational security standards and practices?

16.5 EXERCISES

1. Which of the following US government agencies is responsible for maritime security?
   a. USCG
   b. TSA
   c. DHS
   d. CBP
   e. UN
2. The father of containerization of the global supply chain is:
   a. Malcom McLean
   b. Thomas Friedman
   c. President Eisenhower
   d. President Reagan
   e. Warren Buffet
3. Rapid growth in size and number of ports is due to:
   a. Global tilt
   b. Containerization
   c. Wealth–trade correlation
   d. Comparative advantage
   e. All of the above
4. Which one of the following is an international security standard?
   a. MSRAM
   b. PROTECT
   c. CSI
   d. ISPS
   e. None of the above
5. The source of self-organization in the global supply chain is:
   a. Efficiency
   b. Time
   c. International law
   d. Monopoly power
   e. Large ports
6. The largest port in the world as of 2013 is:
   a. Los Angeles–Long Beach
   b. New York–New Jersey
   c. Seattle–Tacoma
   d. Antwerp
   e. Shanghai
7. The largest chokepoint (betweenness) route in the global supply chain is:
   a. Antwerp
   b. Panama Canal
   c. Los Angeles–Long Beach
   d. Singapore
   e. Suez Canal
8. The WTW (World Trade Web) of 93 trading nations is scale-free and subject to economic disruption on a global scale when:
   a. A highly connected country reduces trade.
   b. A minimally connected country reduces trade.
   c. A highly connected country reduces exports.
   d. A minimally connected country increases imports.
   e. Two highly connected countries default on their loans.
9. MSRAM is a USCG tool that:
   a. Calculates vulnerability
   b. Schedules random patrols
   c. Uses Stackelberg competition to allocate resources
   d. Assesses risk in ports
   e. Prevents terrorist attacks
10. PROTECT is a tool that:
    a. Calculates vulnerability
    b. Schedules random patrols
    c. Uses Stackelberg competition to allocate resources
    d. Assesses risk in ports
    e. Prevents terrorist attacks

16.6 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. Explain why an economic contagion does not equally impact all countries in the world trade network.

B. Why is betweenness a suitable measure of importance in the world trade network? Why is it not the only important critical factor in the network?

C. Containers are getting larger, and so are ports. Is this a symptom of self-organization, and if so, what does it say about the future of international trade? Where will the largest ports be in the future?

D. The United States does not have the largest ports in the world and it is unlikely to ever possess the largest ports. However, the United States has many medium- to small-sized ports located around the entire mainland. Is this beneficial in terms of resilience or not? Explain your answer using complexity theory tools.

E. MSRAM uses ranking instead of optimization to decide how to allocate resources to ports. Is rank order a good or poor strategy? Explain why.

REFERENCES

[1] Friedman, T. L. *The World Is Flat 3.0: A Brief History of the Twenty-first Century*, New York: Picador, 2007, pp. 672.

[2] Transport Research Board of the National Academics, National Research Council. The Intermodal Container Era: History, Security, and Trends, *TR News*, No. 246, September–October 2006.

[3] Cudahy, B. J. The Containership Revolution: Malcom McLean's 1956 Innovation Goes Global. *TR News*, No. 246, September–October 2006.

[4] Sabrin, K. M. and Dovrolis, C. The Hourglass Effect in Hierarchical Dependency Networks, *Network Science*, 5, 4, 2017, pp. 490–528.

[5] Lewis, T. G. *The Book of Extremes: Why the 21st Century Isn't Like the 20th Century*, Cham: Springer, 2014.

[6] Angeles Serrano, M and Boguna, M. Topology of the World Trade Web, *Physical Review E*, 68, 2003, pp. 015101.

[7] Downs, B. The Maritime Security Risk Analysis Model, *MSRAM Proceedings Archives*, 64, 1, Spring 2007, pp. 36–38.

[8] Cowen, D. Container Insecurity: Logistic Space, US Port Cities, and the "War on Terror", in *Disrupted Cities*, ed. S. Graham, New York: Routledge Inc., 2010, pp. 82–83.

17

BANKING AND FINANCE

Wealth is measured in terms of present assets and future productivity, while money is a claim to wealth—not wealth itself. Money is a commodity that flows through a financial system for trading assets such as currencies, securities, equities, and commodities. Currencies simplify and reduce the friction inherent in trading, but it is the financial system that is the critical infrastructure. It is also a dynamical system exhibiting properties of a complex CIKR—fragility, nonlinearity, interconnectivity, and, oftentimes, random chaos.

Money is a particularly important and complex commodity. It has no intrinsic value but, instead, represents wealth in the form of purchasing power—now and in the future. Failures in a financial system are largely failures of confidence in the purchasing power of a currency or assets of value. Present and future asset value establishes a currency's purchasing power. When governments print money "out of thin air" and use it to purchase debt that is paid off sometime in the future, they undermine both present and future value. The economic collapse of 2008–2009 illustrates the threat posed by fiat money and government's willingness to print money. The 2008–2009 Great Recession led to the creation of cryptocurrencies such as bitcoin. Bitcoin and thousands of other cryptocurrencies were developed to avoid central control of money by irresponsible governments. The invention of cryptocurrencies was a direct response to the imbalance between productivity and money.

Productivity is the ultimate source of wealth, while money is a commodity manufactured by the government. When productivity and money supply get out of balance, financial systems become unstable, and the dynamics of an economy become nonlinear. When the imbalance is extreme, severe economic consequences result. It is the job of central banks to stabilize an economy by balancing the money supply with productivity. But it is also the job of speculators and investors to seek a balance between rationality and irrational exuberance that may otherwise lead to major crashes. Trillions of dollars of "wealth" are lost each time a financial bubble bursts and confidence is shaken. What is the source of these threats to the financial system? The fractal market hypothesis (FMH) claims that markets obey long-tailed exceedence probabilities with a relatively high likelihood that they will eventually crash. The FMH further postulates that the cause of financial collapses such as the stock market crashes of 1987, 2000, and 2008 is self-organized crowd behavior—herd mentality—that ultimately grips investors. Herd mentality is a form of self-organization that topples markets and ruins economies. The FMH says that market indexes such as the S&P 500 are fractals that exhibit Levy walks, rather than random walks. Therefore, collapse is intrinsic to free markets. The question is, when will the next crash occur and how bad will it be? The FMH provides tools for estimating when the inevitable collapse is likely to happen next.

Banks are critical components of local, national, and global financial systems, because banks provide the distribution channel between government printing presses and the labor of a productive nation. Accordingly, banks are nodes in a complex financial system including savings and loan companies and investment and stock market companies. Additionally, banks enforce the rules and regulations for

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

exchanging money and the underlying confidence needed for people to trust the essential financial transactions. A financial system's resilience is directly linked to resilience of banks. And bank resilience is directly related to confidence in banking and the present and future value of a currency. This connection is even more significant in the Internet Age, because all virtual transactions ultimately end at a physical bank. For this reason, this chapter focuses on financial systems and banking systems.

The following systems and concepts are surveyed in this chapter:

- *Central banks*: Nearly every country has a central bank for managing its currency. The Federal Reserve System and the Federal Open Market Committee (FOMC) serve as the US central bank and manage the economy through "tools" such as setting interest rates and printing money.
- *The Fed*: The Federal Reserve—the Fed—was established by the Federal Reserve Act (FRA) of 1913. It establishes an interface between the US Treasury and private banks through 12 regional reserve banks and 24 branches. This financial system is connected by the Federal Reserve Wire Network (FedWire) electronic network. Funds are electronically distributed to the banking system through FedWire transactions on a 24 h cycle.
- *Balance sheet*: The Fed maintains a balance sheet containing credits and liabilities. Credits are essentially loans to banks, while liabilities are essentially positive balances such as taxes collected on behalf of the Treasury. The Fed can stimulate the economy by expanding its balance sheet (increasing credits) and contract the economy by contracting its balance sheet (decreasing credits).
- *Printing money*: The US Treasury prints money and sells it to the Fed for distribution. Thus, the Fed stabilizes the economy by controlling the money supply. An excess of approximately 5% more money is printed to replace old bills lost due to wear and damage. If productivity expands more than 5% or less than 5%, an imbalance results that the Fed must accommodate by changing interest rates or controlling the distribution of new dollars and cents.
- *Financial networks*: FedWire is the electronic network connecting the central bank to reserve banks and ultimately the entire banking system. It has a long history going all the way back to telegraphy and Morse code. Currently, it is rushing headlong toward TCP/IP "private over public" network infrastructure that is as vulnerable to malware and disreputable actors as any e-commerce Web site.
- *TARGET*: The Trans-European Automated Real-Time Gross Settlement Express Transfer System (TARGET) and the European Central Bank (ECB) are the equivalent components of Europe's financial system. However, TARGET is subject to political whim compared with the US Fed, because the European reserve banks are aligned with independent sovereign nations instead of a federation of geographical regions.
- *SWIFT*: Society for Worldwide Interbank Financial Telecommunication (SWIFT) is a highly secure member-owned consortium that operates the *SWIFTNet* payment network. This international financial network handles financial flows across national boundaries. In 2012 it linked the banking systems of 212 countries together into one global network.
- *Credit card networks*: Credit card companies like VISA and MasterCard run proprietary private networks that link credit card issuing banks to merchants and the reverse. VISANet is the largest with 3.3 billion cards in circulation, 46 million participating merchants, and 15,900 financial institutions transferring $11.0 trillion/year. VISANet and other credit card companies implement the 3-D security protocol to secure Internet e-commerce transactions that use their credit cards online.
- *3-D secure transactions*: The 3-D secure payment system is a credit card payment protocol designed to secure e-commerce transactions. Its name says how it works—it involves three domains: the user/buyer and his or her bank, the merchant and its bank, and the credit card company. The buyer and his or her bank, the merchant and its bank, and the credit card company validate every 3-D transaction. The 3-D protocol is layered because credentials and authentication are validated at each of the three layers defined by the protocol.
- *Cyber banking and bad actors*: The global banking system is at the heart of nefarious cyber exploitation. Interdiction at the financial network level is the most effective method of halting cybercrime, because criminals ultimately transact business with a bank. However, the rise of cryptocurrency such as bitcoin may circumvent this method.
- *Virtual currencies*: Virtual currencies like PayPal, ApplePay, and cryptocurrencies like bitcoin have established their own financial networks on top of the Internet to secure online transactions or bypass central banking networks such as FedWire and private networks such as SWIFT. PayPal is an intermediary between consumer and merchant that handles electronic payment for the consumer and merchant. ApplePay is a token-based electronic payment system that reduces friction in online transactions by complementing the existing electronic banking system. Bitcoin and other cryptocurrencies replace fiat currency with secure

transactions stored and verified by a distributed ledger called a blockchain. Cryptocurrencies are purposely designed to circumvent official and government-sanctioned banks.

- *Hot money*: When imbalances between money supply and productivity reach extreme levels in one country, money flows to other countries with stronger currencies as determined by interest rates. Speculators borrow low-interest money from the unstable country and invest it in high-interest money countries. This is called *hot money*, because it moves fast and exploits exchange rates as well as productive countries. But hot money has a boomerang effect throughout the world. It tends to entrap nations in liquidity traps and destabilize the financial systems of productive and emerging countries.
- *Dangerous policies*: Central banks risk entire economies using dangerous policies such as "quantitative easing" or rapid expansion of their balance sheets, politically motivated "easy money" to stimulate local economies, and manipulation of exchange rates, because macroeconomics is highly nonlinear. When critical points are reached—such as exceeding economic carrying capacity—entire economies can collapse. This is the largest threat facing many financial systems across the globe.
- *Fractal markets*: Investors are not entirely rational, which means they are subject to self-organization in the form of "groupthink," a type of herd mentality that occasionally grips the investment community. Episodic groupthink replaces "random walk" investment behavior with "biased random walk" behavior that occasionally transforms large groups of bulls into bears or the reverse. Market indexes are subject to Levy walks and therefore are fractals. Groupthink is the source of self-organization in these fractal free markets. Accordingly, the fractal dimension of a market index such as the S&P 500 tells us when and how big the next collapse is likely to be if we can measure and monitor the build up to self-organized criticality.

17.1 THE FINANCIAL SYSTEM

The US banking and financial system is a federation of departments, government bureaus, depositor and investment banks, and committees as shown in Figure 17.1. This complex CIKR has evolved through a series of punctuated events from the first bank of the US, which was established in 1791 and disbanded in 1811, restarted as the second central bank in 1816, and disbanded again in 1836, to the third and current central bank established as the Federal Reserve System of banks by the FRA of 1913. The Federal Reserve System (the Fed) is actually a collection of banks under the control of several boards empowered to operate the banking and financial system on behalf of the US government. It is a central bank with many participating reserve bank branches distributed throughout the country (see Fig. 17.2). The Fed's objective is to operate the system, either expand or contract its balance sheet—a ledger or book containing a list of assets and liabilities of the Fed—and set monetary policy for the nation.

The structure of the Fed and its relationship with the US Treasury (the Treasury) is shown in Figures 17.1 and 17.2. Briefly, the Treasury collects taxes, prints currency, and borrows money. The Fed regulates banks and financial institutions, sets monetary policy, maximizes employment, and attempts to stabilize the economy using a variety of tools such as setting target rates for interest paid on loans. Since the financial crisis of 2008, the Fed has accelerated its purchase of notes and derivatives such as mortgage-backed securities (MBS) in a vigorous effort to stabilize a very shaky economy. The extreme measures taken by the Fed are a test of the agility and resilience of this CIKR.

The principal component of the Fed is its FOMC led by the Chairperson of the Federal Reserve. FOMC is focused on monetary policy, which means the committee determines the amount of money in the system (printing), the interest rates charged on short- and long-term loans, maximizing employment, and stabilizing the economy. Monetary policy has a concrete effect on the FOMC's *balance sheet*, which holds assets and liabilities of the central bank. The Fed buys and sells assets as part of its responsibility to stabilize the economy—an activity that became controversial after the balance sheet expanded dramatically following the 2008 meltdown.

The Fed exerts regulatory and financial control through the 12 Federal Reserve banks shown in Figure 17.1. The New York Reserve Bank in downtown Manhattan is by far the largest bank with over 50% of all assets listed on the balance sheet. It is the central hub of the banking system network. Its proximity to the New York Stock Exchange and other major financial assets such as the "too big to fail" banks makes it an attractive target for terrorists and criminals.

The US Department of Treasury has an account with the Fed just like any other client. This account is where the federal government keeps its tax revenues and also where Treasury assets such as bonds are kept. The US Mint manufactures coins, and the Bureau of Engraving and Printing manufactures bills for sale—at cost—to the Fed. This currency is sent to the reserve banks to stockpile their reserve accounts and fund loans to depositors. The Fed charges an interest rate on these funds. Any "profit" made by the Fed is returned to the US Treasury. Thus, there is a symbiotic relationship between the Treasury and the Fed.

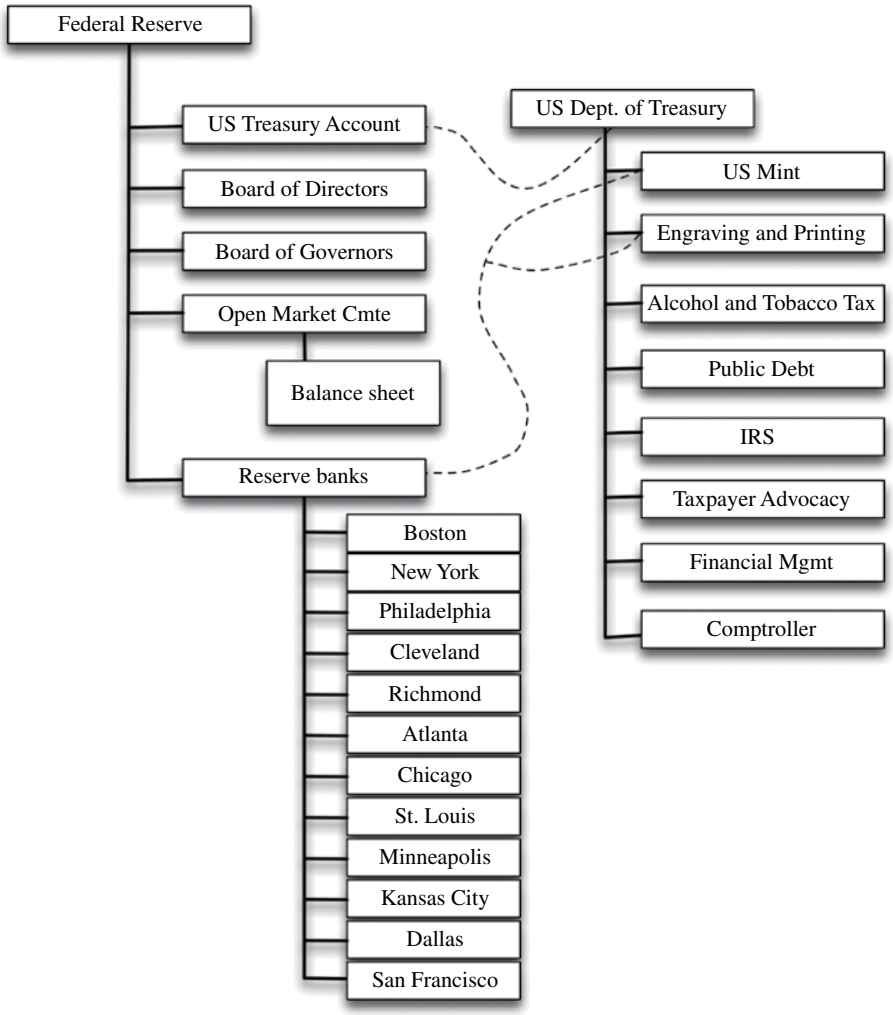


FIGURE 17.1 The Federal Reserve (Fed) and US Department of Treasury (Treasury) are symbiotically related: the Fed sets policy and operates the banking and financial sector, and the Treasury collects taxes, prints money, and issues debt as an account holder with the Fed.

The following is an overly simple explanation of the financial system of the United States. It is intended to provide a basic understanding of money, macroeconomics, and the US banking system so the reader can analyze the US financial system and its vulnerabilities. Each function is illustrated using the hypothetical example below.

17.1.1 Federal Reserve vs. US Treasury

The US Treasury is a customer of the Federal Reserve. The Fed manages America's bank account. Taxes go into the account, and bills are paid from the account. When the US government needs to borrow money, the Fed (and others) buys US Treasury bonds—T-bonds, T-bills, and so on (see Fig. 17.2). These transactions, which typically occur every Thursday at 4:30 P.M., are tracked by posting them on the Fed's balance sheet of assets and liabilities. Typical assets are gold certificates, US Treasury bonds, MBS, and repurchase agreements. Typical liabilities are Federal Reserve promissory notes, US Treasury deposits, and foreign currency. The balance sheet is *expanded* when the Fed buys assets and *contracted* when it sells assets.

Figure 17.2 is a very rough approximation of the money flows in the system. The Treasury prints money, but the Fed distributes it to banks through its 12-reserve bank distribution network. Reserve banks also have branch offices located throughout the country. Financial transactions are very simple—deposits and withdrawals by account holders accumulate throughout the day and are "cleared" overnight by electronic transfers among banks. Banks pay short-term interest on daylight accounts and request more money from the nearest reserve bank when they exhaust their reserves. Banks are no different than individual depositors (consumers) in the way money is deposited and drawn on an account and borrowed when an account reaches zero.

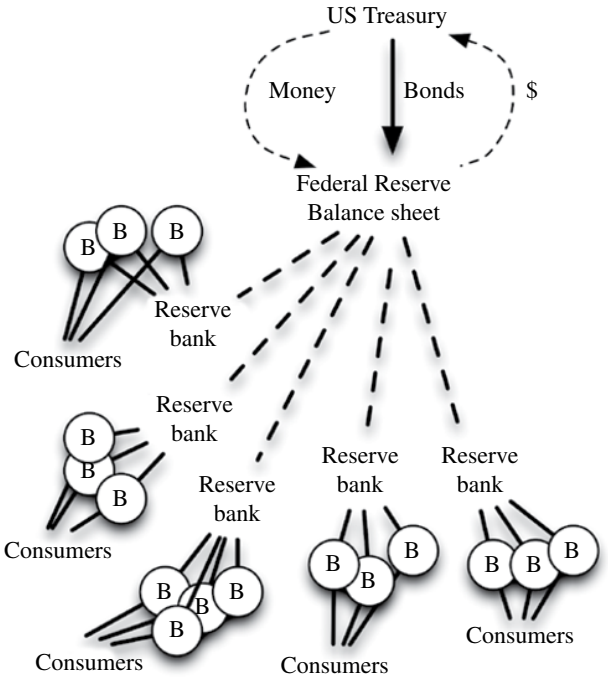


FIGURE 17.2 Money flows from the Treasury to the Fed and then on to the Federal Reserve banks that lend it to other banks.

17.1.2 Operating the System

The primary function of the Fed is to operate the banking system. In general, this means running a bank—the *central bank*—like any other bank, but with one important difference: the Fed can print money. In fact, the Fed *must* print money for the financial system to work. How else is productivity—the ingenuity and hard work of individuals—rewarded? Conversion of productivity into symbolic dollars and cents is an act of wealth creation—the underlying mechanism of capitalism.

Imagine a hypothetical widget manufacturing business—J & S Inc.—run by John and Sally and a hypothetical consumer, Fred, who buys widgets from J & S Inc. The financial transaction between Fred and J & S is tracked through the banking system all the way to the Federal Reserve to illustrate how the system works. Of course the Federal Reserve does not deal with individuals, so this hypothetical example is an exaggeration.

John and Sally formed J & S Inc. to manufacture and produce their patented creation—innovative widgets. John and Sally represent *production* in the form of ingenuity and hard work. They depend on the banking system to translate their ingenuity and hard work into hard cash. In its simplest terms, the Fed trades dollars for productivity, and John and Sally trade their productivity for dollars.

Consumer Fred buys J & S widgets using a credit card issued by his bank. Each time Fred uses his credit card, he is borrowing money from his bank. Therefore, his bank must cover the loan by either dipping into its reserves, borrowing from other banks, or borrowing from the Federal Reserve Bank in its district. One way or the other, Fred's debt must be converted into cash somewhere along the chain of transactions, leading all the way to the central bank. Curiously, by incurring debt, Fred is unconsciously creating wealth, because Fred's purchase rewards John and Sally by converting their productivity into dollars. But this conversion travels a long distance from J & S to Fred and his bank, the banking system, and then back to John and Sally in the form of cash.

Where does the cash come from? Banks borrow money from the Federal Reserve Bank, which in turn buys paper dollar bills and coins at cost from the Bureau of Engraving and Printing and US Mint run by the US Department of Treasury. These currencies are loaned to banks at interest rates set by the Fed. Approximately 10% must be held in reserve by the bank, while the other 90% is used to fund Fred's credit card purchases and make other loans such as home mortgages. Fred's bank can borrow from other banks or from the Federal Reserve Bank at a rate set by the Fed to cover Fred's purchase.

Of course Fred must ultimately pay his credit card bill. Meanwhile, however, the banking system must accommodate the float created by millions of consumers incurring debt like Fred. This float is handled in a number of ways—some short term and other longer term. And of course, banks make a small profit each time monies are transferred, because of interest charges. If these charges are too high, the economy slows down. If they are too low, the economy speeds up. If the economy runs faster than productivity—as measured by *gross domestic product* (GDP)—the economy overheats and becomes unstable. Conversely, if the economy runs too slow, GDP drops and everyone suffers.

17.1.3 Balancing the Balance Sheet

The second major function of the Federal Reserve is to stabilize the economy using Goldilocks economics—not too much money and not too little, but just the right amount. A stable economy is one that stays slightly behind productivity so that the economy does not overheat and slightly ahead of the economy so that the economy does not stall. In practice, this is highly contradictory and nearly impossible to do all of the time.

By expanding its balance sheet, the Federal Reserve stimulates the economy and hopefully stimulates greater productivity. The idea is that cheap and easy money translates into more innovation and hard work. But this approach has limitations as described in Section 17.5. (A liquidity trap may occur where further stimulation has no effect on productivity.)

Stimulation typically means reducing interest rates, which makes it easier and cheaper for companies to borrow, which means jobs are created and people hired. This "trickle-down"

effect increases industrial output—GDP—an aggregate measure of productivity. Trickle down also has a desirable side effect—the multiplier effect that is the result of money changing hands. (Recall that when money changes hands, it is an act of wealth creation, because productivity is exchanged for dollars.)

By contracting the balance sheet, the Fed attempts to slow the economy by elevating interest rates. Because the Fed's target rate increases the cost of borrowing money, it acts like economic friction—putting the brakes on economic activity. It costs banks more to move money around, and consumers pay more for houses, cars, furniture, cell phones, and vacations. Friction can be so onerous that people and institutions stop loaning, entirely. This "credit crunch" is what happened immediately after the stock market plunged in 2008.

This is where monetary policy comes in. It appears that printing money creates wealth and charging too much for it destroys wealth. But money is not wealth, because of buying power. An *elastic* relationship exists between money and productivity that is overly simplified here as follows. Let the relationship between wealth and money be expressed by the equation: $M \sim PQ$, where M is money supply, P is price (buying power), and Q is productivity. This equation says that the amount of money in circulation is proportional to the amount of productivity. Price P is the constant of proportionality that balances money supply and productivity. P is a constant that changes according to monetary policy set by the Fed.

To see how the Fed reasons about monetary policy, consider the rearranged equation, $P \sim M/Q$. Now, prices rise if M rises and fall if Q rises. That is, too much money (M) relative to productivity inflates prices, and too much productivity relative to money deflates prices. Generally, the Fed prints money to increase M and adjusts interest rates to stimulate productivity. After all, if interest rates are low, more businesses will be created or expanded, resulting in more productivity. A low interest rate policy is sometimes called *reflation*, because it is inflationary.

The foregoing simplifications make a number of erroneous assumptions. The first assumption that rarely holds is that the relationship between money supply and interest rates, prices, and productivity is smooth and linear. Nothing could be further from the truth. Instead macroeconomic systems are subject to self-organized criticality, which manifests as tipping points. The dotcom crash of 2000 and the housing market crash of 2008 are textbook examples of reaching and exceeding a sector's tipping point.

17.1.4 Paradox of Enrichment

The Bureau of Engraving and Printing manufactures 38 million bills/day worth $500–$1000 million. Ninety-five percent of these notes are used to replace worn out bills, but the remaining 5% is expansionary. An easy money policy produces too much money and leads to a paradox of enrichment. Recall that the paradox of enrichment says there is a tipping point whereby the ecosystem—the economy in this case—collapses when the predator–prey balance exceeds a certain carrying capacity. Printing too much money enriches the economy until it reaches and exceeds the carrying capacity, which ends in economic collapse. This is a paradox, because too much money leads to too little money!

Low interest rates and an abundance of cash enriched homebuyers for a short time during an expansionary period from 2000 to 2007, but when home ownership reached a capacity of approximately 65%, the housing ecosystem faltered and then collapsed. In short, the Federal Reserve pumped too much money into the housing sector, driving up prices until the economy could no longer support additional purchases at high prices. The bubble burst when the Federal Reserve balance sheet at the end of 2007 stood at $858 billion. Two years later it was $2240 billion—three times larger. Two more years later it had grown another 50% and was approaching $3.9 billion (see Fig. 17.3).

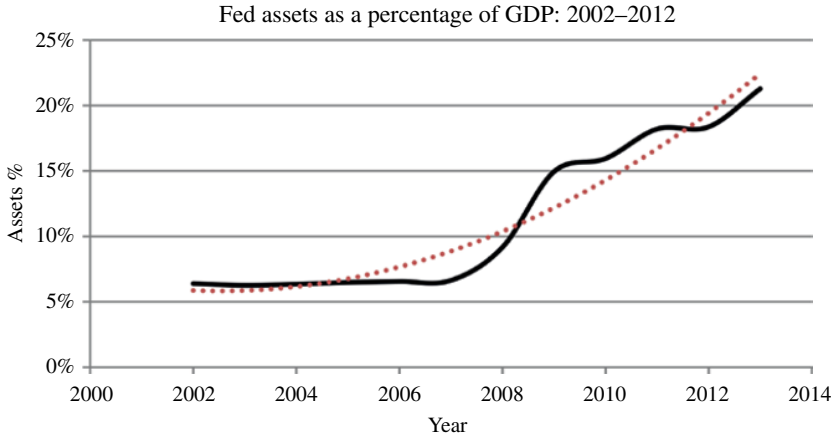


FIGURE 17.3 The Fed balance sheet expanded at an alarming rate following the 2008 financial meltdown. Assets held by the Fed—as a percentage of GDP—grew at an exponential rate.

The steep growth in the Fed's balance sheet since 2008 is historic. As of this writing it stands in excess of 20% of GDP. This matters because the purchasing power of the dollar is only as good as the faith in the US economy. If confidence wanes because investors no longer believe that US productivity will eventually rectify the imbalance, buying power in the equation $P \sim M/Q$ will plunge. This would lead to rampant inflation, and inflation makes it even more difficult to retire debt. Hence debt would beget more debt, in a spiral with disastrous consequences. This is the number one threat to the US financial system, in the modern era.

Clearly, the paradox of enrichment is a nonlinear forcing function operating on the complex banking and finance sector. Macroeconomics behaves in unexpected nonlinear ways when money supply is artificially enriched (without a corresponding increase in productivity). Bad monetary policy threatens stability of the banking and finance system by upsetting the balance of credit and money supply. In summary,

If $M > Q$ then inflation (rising prices).

If $M < Q$ then deflation (falling prices).

If $M \gg Q$ then instability (paradox of enrichment).

Carrying capacity is the underlying capacity of an economy to sustain its debt. Generally, the exact value of carrying capacity is unknown until it is reached and the ecosystem collapses. Apparently, the US economy of 2000–2008 was capable of supporting 65% homeownership, but this critical point was unknown until reaching 69%. Unfortunately, the consequence was a 20% drop in the equity market and a loss of approximately $6 trillion in wealth.

17.2 FINANCIAL NETWORKS

The global financial system is "held together" by electronic networks used to "wire" money from bank to bank and country to country. Prior to the FRA of 1913, cash and gold were physically moved from bank to bank and country to country. Physical movement had obvious risks—particularly robbery and fraud, but it also incurred a cost. Gold is heavy and cash is perishable.

The FRA authorized the Federal Reserve banks to build and operate an electronic funds transfer system that replaced the risky and costly physical transfer system. By 1918, Morse-coded transfers were being wired from bank to bank to settle accounts. The *FedWire* connects the 12 US Federal Reserve banks, their 24 branches, and an additional 7500 other government agencies and foreign banks to the central bank.

17.2.1 FedWire

FedWire electronically connects the US banking system together, but it does none of the transaction processing associated with deposits and withdrawals. It is pure infrastructure. Clearing House Interbank Payments System (*CHIPS*) is a privately held company that does transaction processing and "bookkeeping" typically associated with financial records. Prior to the 1990s private banks implemented proprietary "bookkeeping" software systems to perform accounting and serve customers. Over time, the Fed has promoted greater levels of standardization and resiliency to the system by requiring more standard software.

In the 1960s FedWire consisted of a mainframe data processing center in Culpeper, Virginia, connected to banks by leased telephone lines. But this system was vulnerable to single failures of the privately licensed telecommunications links between reserve banks and Culpeper. The rise of interstate banking in the 1980s forced greater standardization of software and more resilient communication connectivity. The 1990s saw ever more self-organization as networks, software, and servers became standardized and centralized. The 2000s continued this trend as TCP/IP and other commodity technologies were deployed. Today, the FedWire system is a monoculture highly susceptible to the same malware attacks that threaten e-commerce.

FedWire is extremely simple, because it works like a credit–debit checking account. Funds are deposited in an account, and payment made by withdrawing from the designated account. Accounts are balanced at the end of each day, using a *daylight overdraft* system. The *net debit cap* is the amount of daylight overdraft allowed in one business day. As transactions occur during the business day, the balance of deposits and withdrawals changes. Surpluses and shortages must be rectified by the next business day, but in the meantime, the bank is either long or short.

The net debit cap is erased at the end of each business day by transferring funds via FedWire. FedWire transfers contain the names of sending/receiving banks, names and numbers of sending/receiving accounts, and the amount of transfer. The bank pays a duty on daylight overdrafts (the Fed funds rate) and may be required to put up collateral to secure the net debit cap amount. Banks are also required to maintain a reserve as a cushion against major imbalances. As it turns out, making private banks liable for the net debit cap is one of the major reasons that US banks work so well. It is a hedge against the moral hazard of privatizing profits while making the public pay for losses.

The Fed essentially guarantees liquidity, so banks can theoretically never run out of money, but banks must pay a fee. After the 2008 financial crisis, reserves were generally increased to bolster banks against defaults—a critical factor in maintaining confidence in the dollar. This made banks solvent but responsible for the loans. In the end, banks can go out of business if they are poorly run, while consumers are assured that the financial system is safe and secure.

FedWire's principal function is not only to transfer funds between reserve banks, but it is also used to transfer

government securities, collected takes, and other disbursements. Over 500,000 transactions worth $2.7 billion take place every business day. In 2009, 67 participants accounted for 80% of the value of funds transfer. FedWire is essentially the Interstate Highway of banking. Local savings and loan companies and community banks are the streets and alleys.

In many respects the FedWire network is similar to a power grid network. Both systems must balance the flow of a commodity through a self-organized network. Surpluses and shortages of cash are like surpluses and shortages of electrons—something to be eliminated and smoothed out to make sure the system does not collapse from too much or too little of the commodity. Like the power grid, the Fed must provide cash when it is needed to where it is needed on a just-in-time inventory schedule.

17.2.2 TARGET

FedWire is a US network. The TARGET connects European banks together in similar fashion. Together, these two networks account for most of the financial transactions in the world. For example, the euro crisis following the 2008 financial meltdown in the United States was largely a "credit crunch" that occurred because of imbalances in European accounts. European banks were unwilling to loan money to high-risk countries, because they had accumulated too much debt. In 2001, Germany, Netherlands, Luxembourg, and Finland held surpluses, and Italy, Spain, Ireland, Greece, France, Portugal, Belgium, Austria, Slovakia, Cyprus, and Slovenia held deficits in the TARGET system.

The US Federal Reserve System has been criticized for being mostly run by private banks. (The Fed Chairman is a government employee, but the committee is made up of bank executives.) Indeed, the notion of a central bank owned by the government versus private corporations has vacillated between private versus public ownership over the past 200 years. Nonetheless, it appears that private ownership is superior to the public-owned and public-operated ECB and TARGET funds transfer system simply because of private ownership. Moral hazard is the apparent reason for greater resilience in the US system than Europe, because if the privately held reserve banks cannot balance credits and debits, they must sell assets or go out of business. This is not the case with the ECB in Europe. Because ECB banks are aligned with governments, there is no penalty for spending too much money.

Researchers Sinn and Wollmershäuser support this claim:

> An arguably better way to ensure that the TARGET credit ceases to be more attractive for the debtor countries than market credit is the US solution, i.e. the redemption of the TARGET debt by handing over marketable assets to the creditor countries. These could be national government bonds backed by real estate property. As the recipients could sell these bonds in the market and convert them to any sort of preferred assets, the public international credit transfer through the Eurosystem could effectively be avoided. [1]

The threat of going bankrupt and losing personal wealth prevents the US system from overly exuberant expansion to satisfy political objectives. Furthermore, the 12 Federal Reserve banks cut across jurisdictional lines—they overlap political boundaries established by states and metropolitan areas. But in Europe, economic units are equivalent to political units. Therefore, a policy of large public debt in one country tends to drag down the entire European system.

17.2.3 SWIFT

SWIFT is a highly secure member-owned consortium that operates the *SWIFTNet* payment network. Established in Brussels in 1973, it too is pure infrastructure, because it does not maintain accounts with balances nor does it perform clearinghouse functions. It is simply the transmission infrastructure for sending payment orders to 10,000 institutions in 212 countries for processing by other institutions. (The focus on payment orders is significant, because funds flow into accounts—not out of them—a simple but effective way to limit theft.)

SWIFT is standards based, creating and following International Standards Organization (ISO) protocols for sending and receiving payment orders, securing information, and instructing destination institutions on which financial services to perform. In fact, it is a UN-sanctioned standards body for developing secure banking protocol standards. SWIFT has redundant data processing centers, plus one additional center for European transactions. The redundant centers are located in the United States and the Netherlands, and a European-only center is in Switzerland.

The importance of SWIFTNet was demonstrated in 2012 when SWIFT blocked Iranian financial transactions under pressure from European sanctions against Iran. SWIFT claimed that 19 Iranian member banks and 25 financial institutions used the network more than 2 million times that year. This financial blockade is thought to have motivated Iran to suspend its uranium enrichment activities.

17.2.4 Credit Card Networks

Credit card companies like VISA and MasterCard have become a major factor in the economy since Bank of America issued the first credit card in the late 1950s. Initially an extension of the revolving credit line issued by a singly bank, credit cards and the financial networks behind them have become global payments networks that connect consumers, businesses, banks and governments to central banks in nearly every country in the world.

Interestingly, credit card companies do not issue credit cards! Instead, VISANet and other credit card companies provide the network infrastructure for banks, merchants, and consumers to use. Individual banks issue the cards, provide loans, and sign up consumers. VISANet merely completes transactions between merchants and banks for a transaction fee.

The US economy is roughly $15 trillion/year. VISANet—the largest with 38% market share—processes more than $6.9 trillion in global consumer spending/year, equivalent to roughly 50% of the US economy. VISANet can process more than 20,000 transactions per second (150 million/day—200 million at peak loads) using the *3-D security protocol* described in Section 17.2.5. According to its Web site, VISANet links together up to 2.2 billion cards, tens of millions of merchants, 2.0 million ATMs, and 14,600 financial institutions.

Transaction security is especially critical in a large network such as VISANet, so it employs several defense layers to prevent hacks, combat fraud, and protect user's data. PKI technologies described in Chapters 7 and 8 are at the core of these networks, including public key infrastructure (PKI), network intrusion detection (IDS), and artificial intelligence technologies for recognizing malware and hacker attacks. For example, VISA claims, "Our self-correcting network can detect problems in an instant and automatically trigger resolution processes."

The rapid rise of online e-commerce forced a series of new and improved security protocols on the credit card companies. The current widely accepted and used protocol is called 3-D security, because it consists of three layers—the consumer, merchant, and credit card company.

17.2.5 3-D Secure Payment

The 3-D secure payment protocol was designed for online transactions using a credit card. VISA and MasterCard adopted it in the early 2000s to enable Web-based e-commerce. 3-D stands for three domains—the user, the merchant, and the credit card company. Perhaps it should have been called the 5-D payment system, however, because it also relies on cooperation of the user and merchant's banks.

Assuming an online buyer uses his or her credit card to purchase an item from a merchant's Web site, a series of transactions over a secure Internet connection—typically a virtual private network (VPN)—begin when the user completes all purchases and checks out. These transactions involve the user, merchant, and two banks—the merchant and user's banks (see Fig. 17.4):

Step 1. The online cardholder enters his or her name and card number into a Web page provided by the merchant.

Step 2. The merchant validates the user's credit card against the credit card company's directory server.

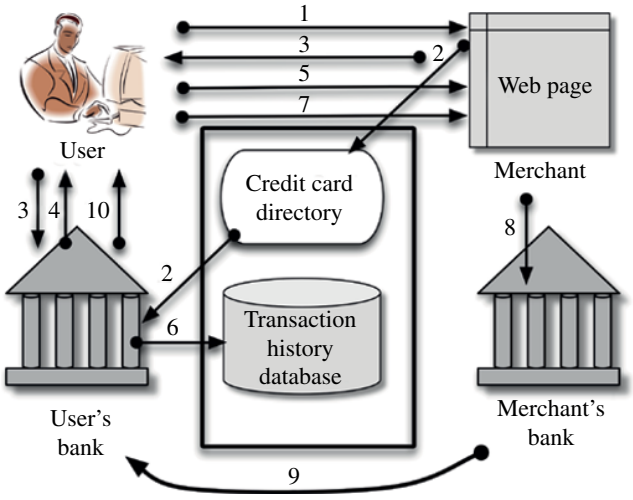


FIGURE 17.4 The 3-D secure payment protocol enables online e-commerce using a credit card such as VISA and MasterCard.

Step 2.1. The credit card company's directory server validates the user's credit card against the user's credit card issuing bank.

Step 2.2. If the user does not have a valid credit card, or the purchase exceeds the card's limit, the transaction is aborted.

Step 3. Merchant server sends card authentication and purchase request information to the user's bank via the user's device—cell phone, tablet, or computer.

Step 4. User's bank receives notice of intent to purchase.

Step 4.1. If user's card is valid, user's bank replies to user's device that purchase is approved.

Step 4.2. If user's card is not valid, the transaction is aborted.

Step 5. User's device replies with authentication information to the merchant's Web page.

Step 6. The user's bank records the transaction information in the credit card transaction history database.

Step 7. The merchant receives secure payer authentication message via the user's device. The merchant's server authenticates the message—typically using digital signatures according to the PKI protocol.

Step 8. The merchant now has all the information and authorization needed to complete the transaction, so the merchant authorizes its bank to complete the purchase transaction.

Step 9. The merchant's bank debits the user's bank with the purchase information. The user's bank makes payment to the merchant's bank, which deposits funds in the merchant's account.

Step 10. The user's bank debits the user's account by the amount of purchase and sends a monthly bill to the user.

This is a lengthy process, but it happens in a few seconds. VISA, for example, is capable of processing 20,000 transactions/second. In addition, it matters little where the consumer, merchant, or banks are located—a feature that makes cross-border hacking attractive.

17.3 VIRTUAL CURRENCY

A virtual currency is any form of currency that works over the Internet. This includes traditional payment systems for electronic transactions such as PayPal and Apple Pay. It also includes cryptocurrencies such as bitcoin, ethereum, and hundreds of others based on a distributed ledger called a blockchain. A distributed blockchain is a ledger copied N times and stored on N nodes in a peer-to-peer (p2p) networks. The purpose of the p2p network is to replace a central bank with a distributed autonomous organization that ensures trust through some kind of protocol such as proof of work (PoW). Such forms of electronic money are called cryptocurrencies because they use cryptographic methods to guarantee authentication, integrity, and confidentiality or transactions.

17.3.1 Intermediary PayPal

One of the earliest online transaction processing systems for secure payment online is PayPal—a middleman between banks and credit card companies and consumers. Online merchants do not want to replace banks, but they want a friction-free banking experience so that consumers can easily and securely buy online products. Intermediary PayPal.com provides the middleman payment processing function as shown in Figure 17.5a.

Using only a login name and password, anyone can pay anyone else by authorizing the transfer of money from one bank account to another. PayPal acts as a third bank to temporarily hold user's money and then make payment to the merchant or an individual at a later time. When used online to buy a product from an e-commerce store, PayPal typically waits for an approval from the user's bank before authorizing payment. PayPal receives a commission for the transactions and also interest on the float.

Assuming user and merchant both have a PayPal account, a user initiates a transaction by emailing a secure message to PayPal.com requesting funds be transferred from the user's bank account to the merchant's bank account. Once payment arrives in the PayPal bank, the merchant is authorized to complete the sale and ship the product. PayPal then transfers the payment from its bank to the merchant's bank.

PayPal broke new ground in terms of electronic payment by allowing anyone to become both buyer and seller. Credit card companies separate buyer and selling into consumer and merchant. A consumer is not a merchant, which means money flows in only one direction. On the other hand, PayPal facilitates money flows in both directions. The intermediary also assumes responsibility for the transactions.

17.3.2 ApplePay

ApplePay introduced a new concept in terms of virtual payment systems. It augments the standard credit card payment system by replacing credit card information with a token and a cryptogram (see Fig. 17.5b). Tokens are 12 randomized digits plus the consumer's 4-digit credit card verification number. The cryptogram is an encryption certificate plus details of the transaction such as product identification, amount purchased, and store information.

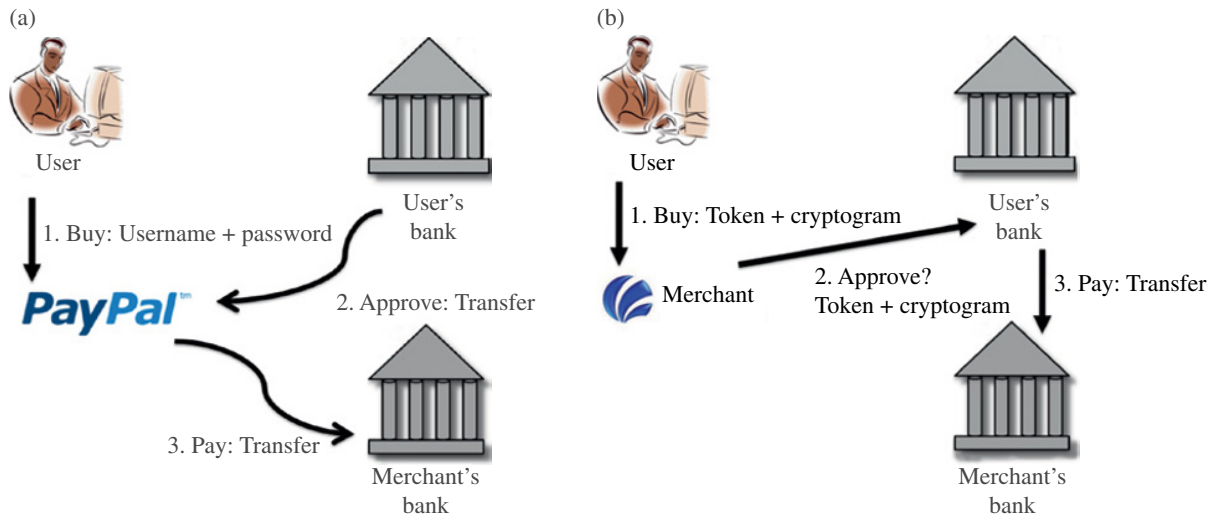


FIGURE 17.5 Two different forms of payment using virtual currency. (a) PayPal intermediary inserts a middleman between consumer and merchant. (b) ApplePay replaces credit card information with a token and cryptogram.

The fundamental innovation of ApplePay is that tokens replace credit card information to prevent man-in-the-middle attacks. If criminals intercept the transaction, they cannot access credit card information. No consumer information is stored in the merchant's point-of-sale terminal, thus avoiding an exploit like the target exfiltration.

Tokens and cryptograms incorporate device identification information so that authorization of payment is done only by the device that initiated the transaction. Furthermore, only the user's bank is capable of decoding the token and cryptogram. In addition, bank transfers take place between banks and not between consumer and merchant. This assures added security because bank transfers are much more secure than consumer transfers.

Assuming users and merchants have registered with ApplePay and banks that process credit cards, the consumer initiates the process by waving his or her smartphone in front of a point-of-sale terminal equipped with near-field communication (NFC). Token and cryptogram are wirelessly transferred to the point-of-sale terminal, which forwards them to the user's credit card bank. Upon approval, an additional transfer of funds from the consumer's bank to the merchant's bank follows.

ApplePay establishes a trusted path from consumer's device to the banking system and ultimately to the merchant's bank account. Authentication is done by face recognition or fingerprint on the user's device. Integrity and confidentiality are guaranteed by the token and cryptogram. Non-repudiation is carried out by the interbanking system that handles payment through bank-to-bank transfers. Tokenization has been copied by a number of other device manufacturers because of its excellent security.

17.3.3 Cryptocurrency

On January 3, 2009, Satoshi Nakamoto began selling a new form of money and operating an associated support system called the bitcoin cryptocurrency system. It is called a bitcoin (BTC), because it is money in the form of bits, and it is called crypto, because the bits are secured by cryptographic methods based on public–private keys. It is still unknown who Nakamoto is, but it may have been Harold Thomas Finney II, a known cryptologist who created a PoW system similar to that used in BTC exchanges and a PGP (Pretty Good Privacy) developer employed by the PGP Corporation. Finney was the first recipient of BTCs in 2009. We may never know who Nakamoto is, because Finney died in 2014.

In his (or her) original paper, Nakamoto defines "an electronic coin as a chain of digital signatures." This is an interesting definition of value because it claims monetary value is more than a token—it is also wrapped up with the unique history of transactions on the token. The value of anything can be defined in terms of the transactions on that something as long as the transactions have not been tampered with and we have a complete record of all transactions. Consider, for example, the title to property such as a house. The title itself has no intrinsic value. Instead, the title represents a certain property that may be valuable or not, depending on previous transactions. When property changes hands, a title company holds the title in escrow while a search is made to identify any liens on the property. If the property has changed hands a number of times, been mortgaged, or been involved in legal disputes, the title leaves a trail of transactions in its wake. This trail is called a chain, and while the title is an integral part of the chain, it is the title's historical record that matters. In the context of digital titles and electronic transactions, an electronic title is a chain of transactions authenticated by digital signatures. Nakamoto's definition makes sense, especially if the title company is replaced by a computer system and the transactions are all electronic. A title, like a coin, is a collection of bits defining ownership and an historical record of all transactions on the title.

The main function of the chain of transactions on a cryptocurrency is to avoid the double-spend problem. This problem arises because a digital token can easily be copied and used repeatedly. When value changes hands, the bits remain. So, how does one prevent additional spending of the same bits? Step one is to keep a ledger of all transactions on the bits, and step two is to distribute trust to a large number of peers that must agree to the validity of the ledger. Trust in the ledger is shared by thousands of peers or nodes called miners, whose job it is to verify the chain of transactions and be paid in cryptocurrency for their trouble.

In addition to a record of transactions copied to thousands of miner nodes, a copy of a user's bitcoin is stored in a user's wallet saved in a user's local device. A BTC in an electronic wallet is a certificate containing encrypted owner information, signed by the owner to authenticate it. These signed certificates, simply called signatures, are based on standard or customized public–private key transactions like any other public–private key exchange process. BTC signatures are cashed in using the owner's private key and purchased using a consumer's public key. To transfer a BTC to Sue, Bob signs it with his private key and uses Sue's public key and an electronic bitcoin exchange to authorize Sue to transfer the BTC to her wallet.

Note the use of an exchange. Unlike the traditional banking system, a cryptocurrency exchange is actually a p2p network of miners working cooperatively to create more wealth by verifying every transaction on the chain of blocks containing transactions. Miners get paid only when they correctly and successfully solve a mathematical riddle that limits the total number of cryptocurrency coins that can be created. This form of scarcity prevents the unlimited printing of money and upward spiral in the price per coin.

The use of p2p networks for sharing information without a central authority traces back to Napster and Gnutella, late

1990s music sharing networks, where digitized music is stored on consumer's personal computers and downloaded to other consumers from wherever the music is stored. The idea of coherent p2p distributed systems goes back to the origins of computer networks, but perhaps the first scholarly study of the coherence problem of a database spread across several computing nodes is due to Lamport *et al*. and the Byzantine generals problem [2]. In simple terms, the problem is to maintain a synchronized and secure database of transactions even though the records are distributed across more than one node. The Byzantine generals problem is complicated due to the potential unreliability of transactions, the lack of a central clock, and the possibility of tampering. When properly synchronized, p2p networks can arrive at a consensus regarding the validity of a transaction. Consensus replaces a central authority as a mechanism of ensuring trust.

The advantage of such p2p networks is the spreading of bandwidth and storage load across many machines and the elimination of a centralized broker or intermediary who might dictate terms or control access. Nakamoto was primarily driven by the desire to eliminate the intermediary—banks, governments, escrow companies, and so on. Intermediaries cannot be trusted, they add cost, and they can reverse transactions. According to Nakamoto, "What is needed is an electronic payment system based on cryptographic proof instead of trust, allowing any two willing parties to transact directly with each other without the need for a trusted third party." But replacing trusted third parties raises a new set of questions:

1. Who or what replaces the intermediary?
2. How can transactions be made tamper-proof?
3. How does a flat or person-to-person exchange prevent double spending?

Nakamoto proposed a p2p network of legers called nodes, as a replacement for intermediaries, a public–private key hash code to prevent tampering, and a clever algorithm called proof of work to prevent double spending. This last challenge—prevention of an electronic coin from being spent more than once by the same owner—was the principal problem addressed by Nakamoto in his foundation paper. Nakamoto asserts, "In this paper, we propose a solution to the double-spending problem using a p2p distributed time-stamp server to generate computational proof of the chronological order of transactions. The system is secure as long as honest nodes collectively control more CPU power than any cooperating group of attacker nodes."

On the surface, the BTC blockchain mechanism appears simple and straightforward. But the servers must do considerable work to verify that the owner is whom he or she claims, the BTC has not been previously spent on the same or different thing, and the ledgers have not been tampered with. In a traditional banking system, bankers prevent double spending and fraud by holding transactions in escrow while the transaction "clears." In the BTC system, verification is performed by miners—operators of the blockchain system that earn BTC for doing PoW verification.

A full explanation of bitcoin is beyond the scope of this book, but transaction processing is easily understood as follows:

1. Init: A user initiates a transaction through an exchange that distributes it to the p2p network of miners.
2. PoW: Each miner attempts to find an encryption key called a *nonce* that is less than a preset value and successfully encodes the user's transaction in preparation to be added to the blockchain. The present value gets smaller over time to limit that total number of bitcoins in circulation and to validate the encryption.
3. Trust: The first miner to find the nonce gets paid in bitcoins and shares the nonce with all other miners in the p2p network. The other miners verify that the nonce works and set about to add the validated transaction to the blockchain.
4. Update: The user is notified and his or her wallet updated with a new balance.

There have been many challenges to Nakamoto's simplistic blockchain architecture in terms of both theory and practice. Theoretically, the p2p network is susceptible to a 51% attack, whereby 51% of the miners collude with one another to falsify entries in the blockchain or simply share the rewards of a successful PoW. In practice, PoW consumes enormous amounts of electrical power, often exceeding the value of bitcoins received. At one point in time, bitcoin miners were consuming more electrical power than the entire nation of Denmark. Bitcoin received much attention in 2012 because of its use by drug dealers and nefarious businesses. Designed to eliminate government intervention and centralized control, by 2019 bitcoin had become dependent on certification by governments and their regulatory agencies.

Cryptocurrencies face even more rigorous scrutiny by consumers. Users must be confidant that their money is secure while stored (in a personal electronic wallet) and when used to complete a transaction. Unfortunately, bitcoin speculation has been rampant, and exchange rates have fluctuated wildly. At press time, it is unclear whether cryptocurrencies will pass from experimental stage to daily use.

Like all modern forms of money, bitcoins are only worth as much as their exchange rate dictates. Virtual currency exchange rates are subject to the same market forces as physical currency. Simply, the value of a currency of any form depends on its scarcity, liquidity, and consumer confidence:

Scarcity: Currency exchange rates depend on many factors, but the principal one is scarcity—how many coins exist versus demand. Some economists claim that

the relatively high exchange rate of gold versus the dollar is because of its scarcity. But even gold expands over time. In 2011 there was an estimated 33,000 metric tons of gold in US reserves with approximately 230 metric tons being added annually. Similarly, bitcoin has an ultimate upper limit of 21 million coins, of which approximately 18 million were in circulation as of 2019. Instead of mining bitcoins from the earth, new bitcoins are generated by solving complex security problems. That is, bitcoin currency "printing" is an algorithm, rather than a printing press. Therefore, in both cases—gold versus bitcoins—scarcity is a factor in determining the value of currency as measured by exchange rates. After 10 years of experimental use, bitcoin exchange rates have fluctuated wildly making it too volatile to be used on a daily basis.

Liquidity: Gold and valuable commodities like silver and oil may hold their value relative to paper money, but they are not very liquid. First, it is difficult to spend gold bars or trade barrels of oil for food. Second, exchanging a commodity for spending cash requires much more infrastructure than buying a cup of coffee with a credit card. Physical money is more mobile and more liquid when it comes to handling transactions. But in the digital world, even physical money like dollar bills is too cumbersome. Thus, bitcoins and other forms of cryptocurrency are far more liquid than physical money when it comes to online transactions. However, bitcoin and other cryptocurrencies have not been entirely friction-free. Bitcoin is too cumbersome for most consumers to use like paper money.

Confidence: Confidence in currency of all forms depends on the willingness of users to accept it. This is true whether money is symbolized by gold bars, dollar bills, or virtual currency. While currencies are often pegged to gold, the reverse argument is also valid—the value of gold can be pegged to the dollar or bitcoin. Money is a symbol regardless of its form. Moreover, the value of money is relative to the value of something else—for example, its exchange rate. If consumers are convinced that a certain virtual currency is worth more than another form of currency, then it is. Exchange rates are mostly established by confidence, not by absolute values. In fact, exchange rate arbitrage is a major business around the globe. Confidence in cryptocurrencies has fluctuated wildly, also, leading to reluctance on the part of consumers to broadly accept cryptocurrency as a means of daily transaction.

Virtual currencies have established strong credentials with respect to scarcity but less with respect to liquidity and confidence in their long-term exchange rates. This uncertainty has led to extreme bursts of speculation in some virtual currencies such as bitcoin. If confidence in the Fed wanes, virtual currencies increase in value. If consumers believe bitcoins will be worth more relative to a government-sponsored currency, bitcoins will increase in value.

Money—in any form—is fragile because it must strike a balance between supply and demand. And because of its symbolic value, currency of any form must ultimately relate to productivity. If the fragile thread linking a currency to productivity is broken, then people quickly lose confidence in the currency. Therefore the largest threat to any financial system is lack of liquidity and loss of confidence in the value of productivity represented by the currency, regardless of what form money takes.

17.4 HACKING THE FINANCIAL NETWORK

As national and international banking and financial institutions adopt TCP/IP and other Internet technologies, their vulnerability to cyber exploits also increases. First, the Internet was designed to be open and lacks fundamental security features like encrypted source and destination addresses. Second, it is a monoculture of identical servers, algorithms, and operating systems—making it relatively easy for terrorists and criminals to exploit banking systems en masse. Fragility is intrinsic to the Internet, and as banks adopt the Internet's protocols, they also adopt its weaknesses.

An interesting and important study of spam operations carried out by researchers led by Stephan Savage at the University of California at San Diego illustrates how disreputable people exploit the financial system using the Internet [3]. The UCSD study focused on spammers, but the techniques used by spammers suggests vulnerabilities in the banking system, itself, because the UCSD team concluded that successful nefarious online activities come down to hacking the global financial network. If spammers can manipulate the banks, then more threatening criminals can also manipulate them.

The UCSD team concluded, "evidence that the payment tier is by far the most concentrated and valuable asset in the spam ecosystem, and one for which there may be a truly effective intervention through public policy action in Western counties." In other words, if you want to stop online crime, the financial system is the best place to focus attention. Furthermore, countermeasures are most effective if defenders focus attention on a handful of banks, rather than the thousands of banks spread throughout the globe.

Figure 17.6 summarizes the complex series of transactions carried out by spammers as they troll the Internet looking for consumers interested in contraband recreational drugs, knockoff brands, and counterfeit software. (The product could be illegal drugs or other illegal products as well.) The series begins with an email generated by *bots* running on *zombie* computers culled from innocent users. A consumer must act on the spam to start the series of steps leading up to purchase of contraband Viagra. This is step 1 in Figure 17.6.

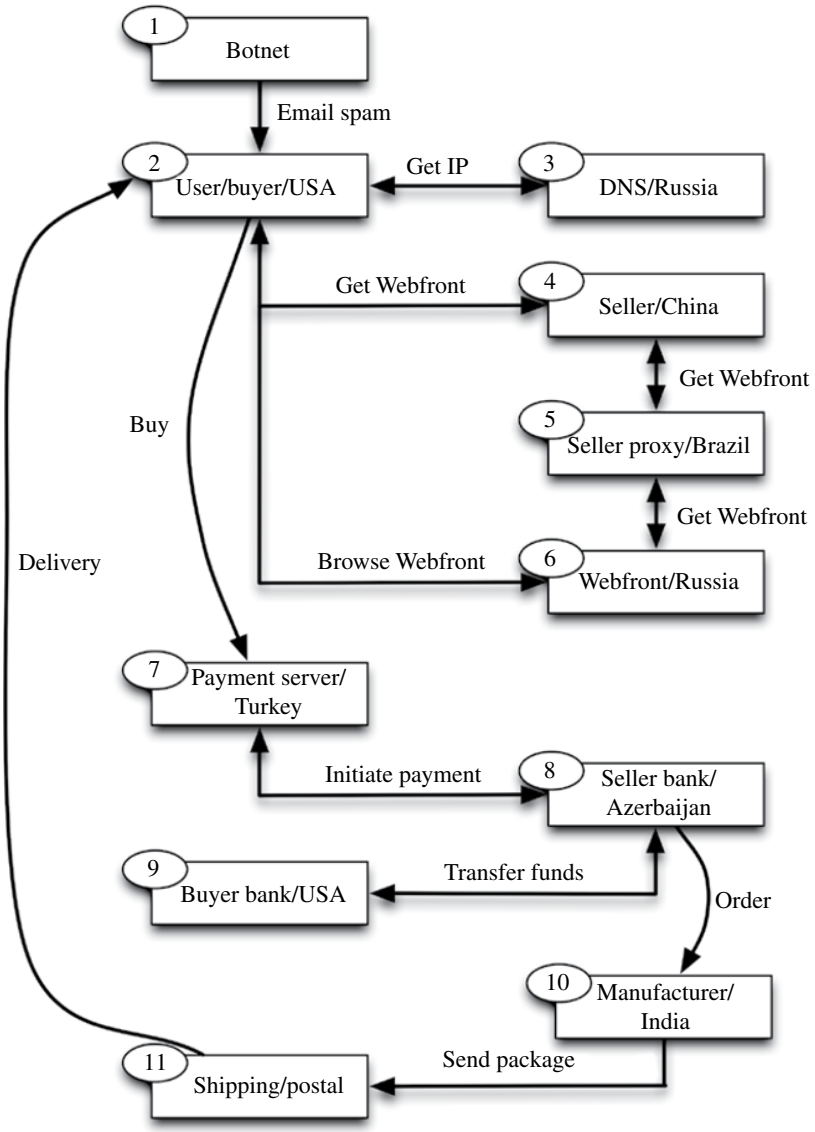


FIGURE 17.6 Transactions in the UCSD study span the globe.

By clicking on a URL embedded in the spam email, a user/buyer in the United States initiates steps 2, 3, 4, 5, and 6. The initial click returns an IP address from a DNS in Russia (Domain Name Server converts a URL like www.mystore.com into an IP address like 100.010.04.011), but the Web page can be located anywhere in the world, perhaps to avoid local laws or blocked IP addresses. In this example, a server in China uses a proxy server in Brazil, which in turn links the user to a Web page back in Russia. The user/buyer does not see these behind-the-scene transactions, however. Instead, the user/buyer sees a Web storefront offering drugs at low prices without requiring a prescription.

The transaction bounces around the globe as a payment server in Turkey takes the buy order and processes it using a seller's bank in Azerbaijan—see steps 7, 8, and 9. The Azerbaijan bank transfers funds from the buyer's US bank and places the drug order with a manufacturer in India—see step 10. The Indian manufacturer receives payment and ships the order to the user/buyer in the United States via a logistics company such as UPS or FedEx.

The transactions take place at lightning speeds, of course, crossing the Atlantic Ocean several times. The only way to stop it is to intervene in the global financial system. Fortunately, there is a very small set of blocking nodes in this network. According to the UCSD study, "… just three banks provide the payment servicing for over 95% of the spam-advertised goods in our study." Preventing these three banks from completing transactions from filtered IP addresses dramatically reduced this criminal activity. Between 2010 and 2012, spam traffic dropped precipitously as a result of actions taken by banks.

17.5 HOT MONEY

The rapid rise of TCP/IP banking exploitation on a global scale may pale in comparison with the threat of *hot money* flows across national borders. Ronald McKinnon and Zhao Liu define hot money as, "speculative money from carry traders flooding into emerging markets with higher interest rates [provoking] domestic inflation [leading] to local currencies being overvalued. When emerging market currency exchange rates are not tied down by official parities, their ongoing appreciation induces more hot money inflows, as one-way bets on currency appreciation are induced" [4]. An imbalance between exchange rates due to radically different interest rates creates hot money flows. These flows can capsize an entire country.

Hot money flows combined with *liquidity traps* as experienced by Japan, and possibly the United States and Europe following the 2008 financial crisis, tend to accelerate economic collapse because of a vicious financial cycle described below. A liquidity trap is a type of *paradox of enrichment* failure where injections of cash into a banking system by the central bank lead to less economic growth rather than more. When low interest rates and excessive printing of money by the Federal Reserve stretch the balance between productivity and currency in circulation beyond its elasticity, inflation spikes and the economy collapses. Liquidity begets less liquidity when people hoard cash, stop purchasing goods and services, or lose confidence in the system due to war or in anticipation of financial shocks.

Hot money speculators take advantage of a central bank's vain attempt to stimulate an economy beyond its *carrying capacity* as bounded by productivity. Central banks like the Federal Reserve are under political pressure to stimulate the economy by printing money, lowering borrowing costs, and expanding their balance sheet. When carried to extreme, this locks entire nations into a vicious cycle as shown in Figure 17.7. The ultimate outcome is collapse of undetermined size and consequence. This cycle began building in 2003 and encompassed the entire world by 2012.

The principal villain in Figure 17.7 is the US Federal Reserve, because the US dollar is the largest currency in circulation and the world's reserve currency. As the Fed tries to stimulate the US economy by lowering interest rates and printing dollars, the currency of emerging economies tend to inflate. The currencies of Brazil, Russia, India, China, and South Africa (BRICS) rocked as exchange rate shocks ripple through their economies. The cycle works as follows:

> The Fed lowers interest rates and prints money, thus upsetting the balance between real productivity and the value of the dollar. This sudden change in the dollar shows up as sudden changes in exchange rates. Inflation makes it attractive for speculators to borrow money in low interest rate dollars and invest it in high-return investments in the BRICS such as China and India. Hot money flows from the US to the BRICS. But this is only the initial consequence of hot money flow.
>
> Cheap and plentiful dollars inflate the price of commodities such as food and energy. Emerging countries suffer more than industrial countries because food is a major

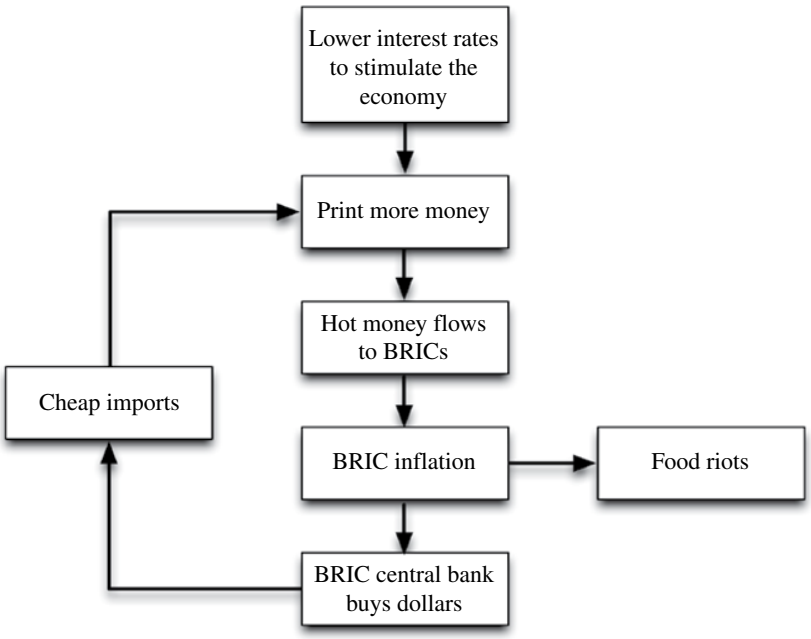


FIGURE 17.7 Hot money flows from low interest rate economies to high interest rate economies in a vicious cycle.

percentage of people's budgets in emerging countries. One consequence was the April Spring, which McKinnon and Liu claim was a direct reaction to the Fed's stimulus policy. "In December 2010, it was a poor Tunisian food vendor that immolated himself—thus starting contagious riots throughout the Arab world. Unfortunately, the Arab Spring (as the name implies) was interpreted by Western diplomats as a sudden longing for democracy and a desire to throw out corrupt dictatorships—and thus it was widely believed that the West should support the rebels. If the Arab Spring had been recognized as mainly a food riot, the response of Western governments would have been more measured in taking sides, while focusing more actively on monetary measures to dampen cycles in primary commodity prices."

BRICS central banks respond by buying dollars to "soak them up," and temper exchange rates. This is why the US debt held by China is so large. Without this policy, China and other BRICS would be swamped by dollars, and their goods and services would inflate, leading to an export slowdown. As described in Chapter 16, the wealth of a nation is strongly correlated with exports. A drop in exports causes a drop in the overall economy.

To keep their economies going, the BRICS continue to export cheap products and services to the US and other dollar-enriched nations. To keep the voting public happy, the Federal Reserve continues to print more money so that voters can buy more cheap imports. To soak up abundant dollars, the BRICS buy them. Thus, the vicious cycle spirals deeper into debt—the liquidity trap that stalled Japan in the 1990s.

The liquidity trap tightens as the Fed lowers interest rates and prints money in a vain attempt to "get ahead" of the cycle. But it is impossible to get ahead of a cycle that the Fed is a major part of. This is why the liquidity trap is a trap. Indeed, the liquidity trap only gets tighter as history has shown.

17.5.1 The Dutch Disease

The *Dutch disease* was explained in an article appearing in *The Economist* in 1977 as the ensuing calamity that occurs when productivity in one part of the world declines and money flows into another part of the world. The name derives from an economic incident that happened in the Netherlands when the manufacturing sector suddenly declined after the discovery of a large natural gas field in 1959 [5]. Discovery of a valuable resource like natural gas should benefit an entire nation, but as it turned out, the discovery triggered a form of enrichment that distorted the Dutch economy through a complex interaction of money flows between the two economies.

The natural gas discovery sharply inflated a sector of the economy, sending commodity prices through the roof. The natural gas sector thrived, driving up labor rates and commodity prices, which in turn, eradicated prosperity in the broader economy. By enriching one sector, the natural gas economy crashed the other sectors.

W. Max Corden and J. Peter Neary developed a general economic theory in 1982 to explain bubbles like the Dutch Disease [6]. In their model, there are two sectors—a *booming sector* and a *lagging sector*. The booming sector is usually based on natural resources like natural gas, oil, gold, copper, or agriculture. The lagging sector generally refers to manufacturing. A surge in the booming sector ultimately increases the demand for labor, which increases salaries in the sector, which then shifts production (and hot money investment) away from the lagging sector.

Dutch disease also contains elements of *Gause's competitive exclusion principle*. As you recall, this principle says only one dominant species can emerge from an ecosystem starting with a field of multiple competitors. In this case the booming and lagging sectors compete for labor. Eventually, the booming sector wins because of its advantage—it pays higher wages.

The Dutch disease example illustrates the deep nonlinearity of macroeconomics and the unintended consequences of enrichment. By printing money and lowering interest rates, the Fed enriches one sector of the global economy while disrupting the economy of BRICS. But BRICS are not the only casualties. The US economy itself is ultimately the largest casualty because of the liquidity trap. Where does it end?

17.6 THE END OF STIMULUS?

Perhaps the most consequential threat–asset pair in the banking and financial sector is the financial system itself. The Fed may not be sophisticated enough to understand the complexity of the system they are attempting to stabilize. Instead of stabilizing it, their policies may collapse it. Thus, the Federal Reserve policies themselves become weapons of mass wealth destruction. How does a central bank get out of its liquidity trap, and what happens when stimulus is removed? Theoretically, slow tapering is the answer, but nonlinearities may overtake even incremental changes—especially when crossing from stimulus to negative stimulus.

Figure 17.8 contains results obtained from a computer model of productivity as measured by GDP calibrated to match the growth of US GDP from 1980 to 2012. The nonlinearities in Figure 17.8a occurred because of the 2008 financial crisis and expansion of the Fed's balance sheet immediately following. Note the close match even when nonlinear effects of enrichment perturb growth. The model uses a combination of regression analysis and predator–prey nonlinearity to approximate actual GDP and its projection, assuming 2.5% annual growth.

Figure 17.8b contains results from running the exact same model with one exception—economic carrying capacity gradually declines after 2012. This gradual decline simulates lower productivity, which implies less economic

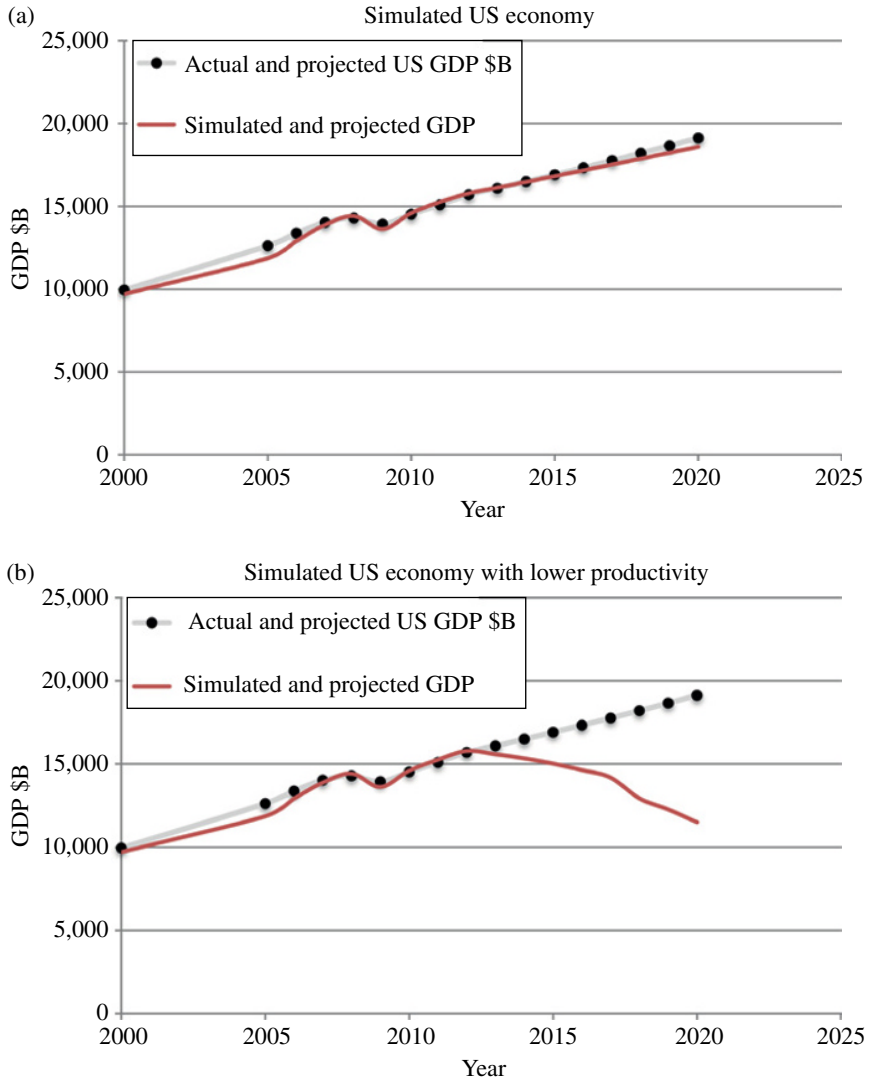


FIGURE 17.8 Simulation of the nonlinear effects of economic expansion on GDP of the United States. (a) Comparison of actual and simulated model of GDP versus time validates the model used in (b). (b) The same model used in (a) is modified by lowering the carrying capacity of the US economy due to lower productivity.

activity. What happens when consumer purchases decline? The projected GDP departs from the 2.5% growth model, simulating the anticipated economic slump following tightening money supply.

It is impossible to predict the future, but assuming that productivity declines by a small amount results in the reversal of economic growth. As productivity declines, so does the economy. But this is a classical model of a much more complex financial system. In the final section, a nonlinear fractal model of the economy is shown to contain the seeds of its own destruction. Using various financial market indexes, such as the S&P 500 and Dow Jones Industrial Average, complexity scientists have shown that free-market systems are intrinsically prone to SOC and, therefore, eventually fail.

17.7 FRACTAL MARKETS

Perhaps the biggest threat to the stability of a national financial system is the classic bubble, as experienced in Minsky moments, meltdowns, Dutch diseases, and so on. These black swan events have become rather common in the modern age, resulting in huge financial consequences. For example, the 2008 financial meltdown eliminated $3–6 trillion in wealth—far more than the consequences of the 9/11 terrorist attack and wars in Iraq and Afghanistan—more than the entire GDP of most other countries. Bubbles affect everyone in the economy.

An important question is, "why are financial systems, such as the free-market system of the US, prone to instabilities?" The answer lies in a deep understanding of complexity

theory, self-organization in social networks, and the fractal nature of the stock market. Since the 1990s, physicists and mathematicians have challenged classical economic theories like the efficient market hypothesis (EMH) on the basis that economic theories have failed to explain reality. These so-called econophysicists are building a new theory of economics based on complexity theory and fractals. Their claim is that financial markets are nonlinear, self-organizing, fractal systems driven by human frailty—principally the common human tendency to be swept away by episodes of herd mentality. Instead of efficient markets, the highly connected world of the twenty-first century is driven by emotion. This leads to rare but disastrous collapses.

17.7.1 Efficient Market Hypothesis (EMH)

The EMH—an economic model pioneered in the 1960s and the model currently employed by the Federal Reserve even today—argues that markets such as the S&P 500, Dow Jones Industrial Average, and NASDAQ are rational and efficient. That is, investors act rationally because they use all available information—typically company reports and news of current events—to make investment decisions. Furthermore, their buy and sell patterns aim to maximize gains and minimize losses. Investors want the best possible returns, which means risk is balanced with reward. The upshot of this rational behavior is a Brownian motion random walk—the number of bulls and bears buying and selling is like the random movement of molecules in a room full of air. Molecules bounce off of walls and one another without bias. They take a random walk from collision to collision in accordance with the normal distribution.

Similarly, market prices traverse a random walk as they "walk from transaction to transaction," in accordance to a normal distribution. Some investors buy, some sell, and others hold. Furthermore, some investors are day-traders, some are long-term traders, and others make buy/sell decisions based on price and current events. When stirred together, the ensemble of investors behaves just like molecules in a room full of air.

Economists of the EMH school, such as Eugene Fama, model price movements as random walks and ignore the underlying human emotion that drives many investment decisions [7]. Buying and selling is simply an optimization problem to be solved by observing the fundamentals of business. If traders buy and sell according to the rules of Brownian motion, then stock fluctuations should vary proportional to the square root of time (or number of trades). For example, if a stock index is priced at \$100 today, after t days of trading, its price should go up or down by an amount proportional to $t^{0.5}$. In fact, a market that behaves in this manner is considered purely random, and prices ultimately regress to the mean. This is why traders use the 200-day average of a stock to determine if its price is too high or low.

EMH, which is an idea over 100 years old, rose to prominence in the 1960s and continues to enjoy mainstream acceptance today. But it began to fail by the 1990s, because it no longer matched reality. EMH could not explain bubbles and crashes—black swan events like the 2008 meltdown were impossible under EMH and yet there they were. This led a radical group of economists to propose the FMH—a model based on complexity theory used throughout this book to explain catastrophes.

17.7.2 Fractal Market Hypothesis (FMH)

The FMH, advanced in the 1980s and 1990s by pioneers such as Edgar Peters and Didier Sornette, discards the random walk idealization in favor of the Levy walk model. Recall that Levy walks are pseudorandom walks that obey a long-tailed power law. Most displacements in distance, consequences, time, or price are small, but some are long. Extreme jumps occur with probability greater than expected by Brownian motion but are comparatively rare. Stock prices, for example, tend to make many small up or down movements, but on rare occasion they make large up or down movements. Even more unusual is the black swan outlier whereby a stock—or entire stock market—suddenly falls by 20, 30, 40, or 50%. The EMH cannot explain this long-tail effect, but FMH can.

Peters was one of the first traders to realize that stock markets are biased random walks that obey a Levy flight pattern rather than a Brownian motion pattern [8]. That is, prices fluctuate according to a long-tailed power law, which means prices form fractal patterns. If a stock index was a busy highway, traffic jams infrequently occur, but when they do, the size and elapsed time between jams mimic a Levy walk. Most of the time traffic flows smoothly with small changes in speed. But when a platoon of cars forms, the entire platoon slows down, and when one car stops, the others jam together and stop, or nearly stop. Traffic flow is episodic, and so is the real-world stock market.

The analogy with traffic is not especially good, but it serves to illustrate how self-organization of both traffic patterns and investment patterns happen. In a financial market, the ebb and flow of bulls and bears produces dynamic self-organizing social networks of traders that tend to buy and sell in smooth patterns until a "traffic jam" forms. The smooth flow of trades reaches a self-organized criticality when trades jam together in lock step. That is, traders become organized and synchronized instead of random and disorganized. They form platoons of buyers that drive prices to unreasonable heights, or they may form a platoon of sellers that crash the market. Their sudden synchronization leads to bubbles and their collapse.

Peters argues that investors act as a herd—the groupthink of a small segment of the investor network spreads like an

epidemic to other members, which in turn increases the spectral radius (self-organization) until SOC is reached at which point a cascade of sell orders happen all at once. This crashes the market. This is caused, in Peter's model, by investors with different investment horizons and objectives—some are day-traders, some are buy and hold, and others are mixed somewhere in between. But when they self-organize and act as a platoon, the market transitions from disorders to ordered or, in complexity theory terms, from random to structured.

Peters proposed a *rescaled range analysis* tool for finding the fractal dimension of specific markets. The rescaled range of price fluctuations obeys a power law typical of fractal behavior, with the Hurst exponent H in place of 0.5, as in the EMH model, where q is the fractal dimension of the market [9]:

$$R/S \sim t^H$$
$$q = 2 - H$$

When applied to the S&P 500 index leading up to the 1987 crash, $H = 0.72$. H always lies between zero and one. What does this mean?

If $H = 0.5$, the market fluctuations are random, but if $H > 0.5$, the market is biased toward Levy walk behavior, instead. If $H < 0.5$, the market is headed for disorder and becomes erratic and falls apart. This has never been observed. Rather, $H = 0.72$ means there is a substantial amount of bias toward short-term trades instead of long-term trades. $H = 0.72$ means the S&P 500 is self-organized (herd mentality) and prone to collapse:

1. $0 \le H \le 0.5$: The financial system underlying the index becomes disorganized and dies out.
2. $0.5 < H \le 1.0$: The financial system underlying the index is self-organized, because there are more bulls than bears or more bears than bulls. We should expect occasional collapses.

Peter's rescaled range analysis tells us whether an index is self-organized, just as spectral radius tells us a network is self-organized. But it does not tell us when SOC will lead to collapse of the financial system. Is there a way to anticipate the inevitable crash?

17.7.3 Predicting Collapse

Didier Sornette, professor of entrepreneurial risk at the ETH Zürich, proposed a method of predicting the collapse of financial bubbles by analyzing the fractal structure of stock indexes such as the S&P 500, Hong Kong, and other exchanges [10]. Sornette argues that self-organization of buyers and sellers leads to a crash when sellers unload their holdings all at once. That is, they act in unison much like a herd of stampeding cattle. Self-organization is an episodic formation of bearish investors whose buy and sell patterns can be detected by careful analysis of fractal dimension and oscillations in prices.

According to Sornette, price levels leading to a crash exhibit log-periodic behavior, leading up to a singularity in the index. In log-periodic behavior, *Sornette waves* are simply oscillations that increase in frequency as prices near collapse. In traffic jam terms, an impending traffic jam 5 miles ahead "telescopes itself" by sending waves upstream. As the waves of stoppage work backward from the point of a stalled car, accident, or slowdown, they increase in amplitude but decrease in frequency. From the point of view of an unwary motorist approaching the jam, waves of congestion increase in frequency and decrease in amplitude as the motorist nears the point of congestion. Similarly, a stock price singularity telescopes the impending event backward in time as follows.

> ***Sornette Waves:*** *The amplitude of price fluctuations decreases, and frequency of price fluctuations increases as the index approaches the point in time of collapse. [The oscillations "blow up."] A singularity appears at the point in time of collapse, because amplitude reaches zero and frequency reaches infinity.*

Figure 17.9 illustrates the log-periodic behavior of the S&P 500 over the 2-year period, leading up to the October 1987 crash. Sornette waves are shown tracking the oscillations in price as time passes. Their amplitude slowly diminishes as frequency increases. At the point of failure, Sornette waves "blow up." A mathematical singularity marks the point of imminent collapse.

Sornette waves are produced by a combination of power law behavior and oscillations based on the logarithm of time to collapse. As time approaches the collapse date, the difference between time and collapse time shrinks to zero, which reduces amplitude according to a power law and increases frequency according to a log-periodic law. At the singularity, the log-periodic component becomes unbounded. This singularity marks the date of the crash.

Sornette used his algorithm to predict the 1987, 2000, and 2008 crashes. Why does Sornette's FMH work? Figure 17.10 shows the result of a simulation performed by the author. Nodes represent investors, and links represent the influence one investor has on another. Nodes are dark if they are bulls and blue if they are bears. One node is permanently painted dark and another is permanently painted shaded. A pair of linked nodes is considered neighbors, and neighbors influence one another by sharing information about the market. As "red information" spreads through the network, it influences neighbors by increasing the likelihood that the

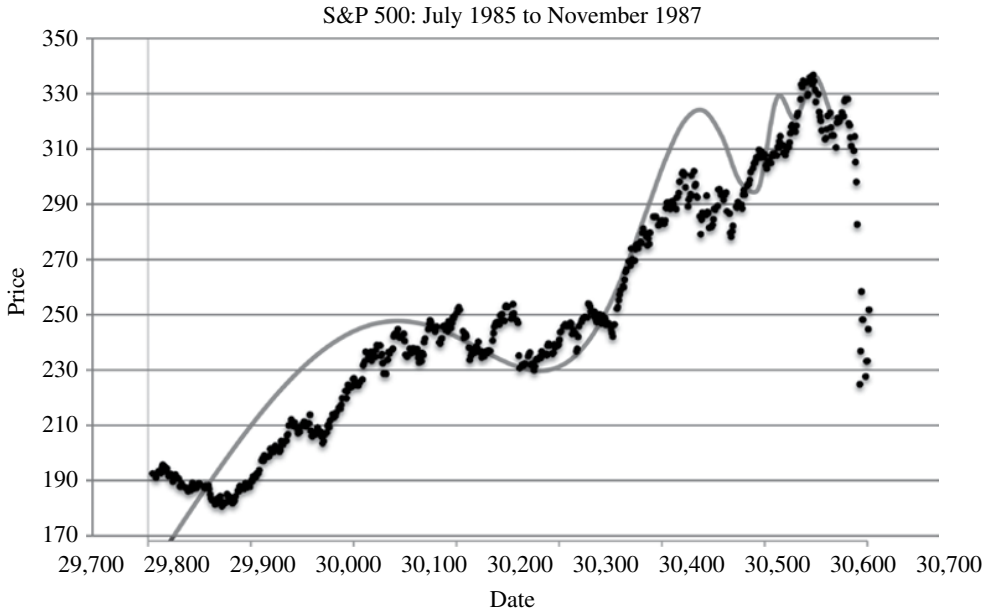


FIGURE 17.9 The S&P 500 index crashed in October 1987. Log-periodic waves are shown as a best-fit curve (solid line) that increase in amplitude and frequency until the log-periodic function blows up.

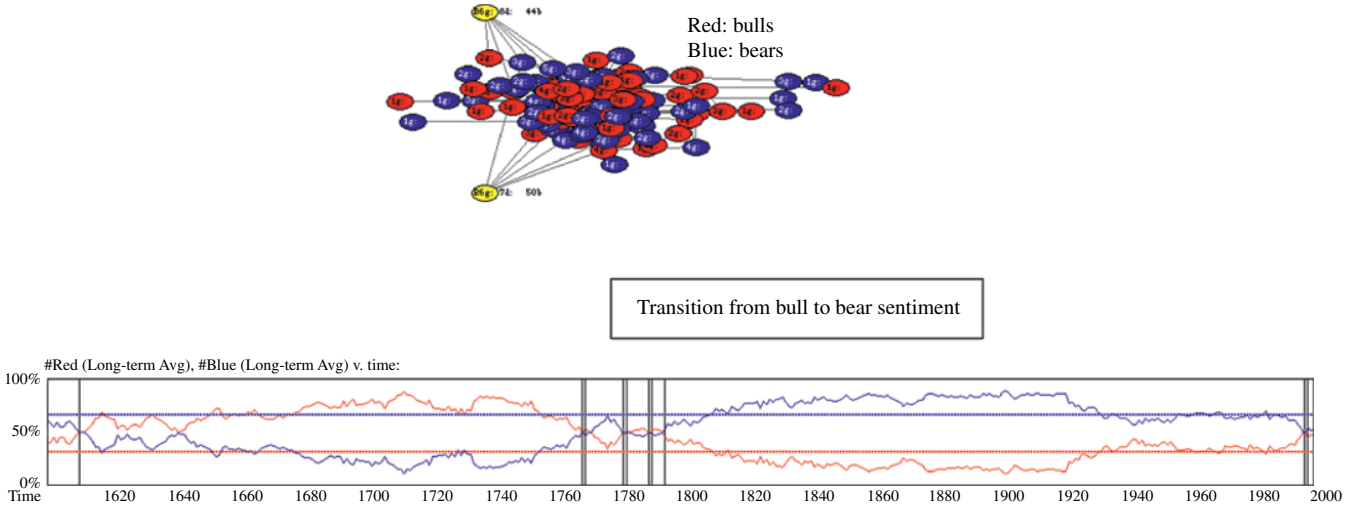


FIGURE 17.10 The S&P 500 index from July 1985 to October 26, 1987, is shown as solid dots and Sornette waves are shown as a solid line. Sornette waves increase in amplitude and frequency until they reach a singularity at the point of collapse.

neighbors will also become dark. Therefore, this simple simulation approximates the spread of human emotion regarding whether to buy or sell.

What happens over time? As dark and shaded influences spread from neighbor to neighbor, sentiment changes—sometimes oscillating between dark and shaded and other times remaining one color for long periods of time. This simulator mimics self-organization of social networks by repeating a very simple rule at every time period and observing the formation of platoons or traffic jams around dark or blue positions [11].

Rule: *The probability of becoming a dark (bull) node is proportional to the number of neighbors that are also dark. Similarly, the probability of becoming a shaded (bear) node is proportional to the number of blue neighbors. Initially one node is dark and another node is shaded.*

The number of bulls (dark) and bears (shaded) oscillates over time. If the system is balanced, the number of buyers equals the number of sellers. In this case, the market behaves like a random walk. But as time passes, the number of bulls dominates the number of bears for a time and then reverses

so that the number of bears (shaded) dominates the number of bulls. A snapshot of these changes in dark/shaded sentiment is shown in Figure 17.10. Note the "chaotic fluctuation" at crossover points where dominance changes from bulls to bears, or the reverse. This choppiness is what causes Sornette waves to appear (and disappear) in the price, indicating an impending "sea change" in sentiment. When overwhelming conviction changes direction, the market dramatically rises or falls.

This simple simulation is only one confirmation of the underlying self-organization that drives markets either up or down. Most significantly, it shows how irrational behavior in the form of a herd mentality can take hold of a financial system. Specifically, this model is ignorant of information or rationality that most people possess. It ignores current events, paradox of enrichment, and other explanations for why sentiment changes. However, it explains why fractal markets intrinsically create bubbles and then burst them. The boom-and-bust cycles of free markets are a direct consequence of their complexity. As observed throughout this book, every complex system contains the seeds of its destruction.

17.8 EXERCISES

1. Which of the following is the US central bank?
   a. TARGET
   b. The Federal Reserve
   c. The US Department of Treasury
   d. The US First National Bank
   e. ECB
2. How is wealth created in the US economy?
   a. The Fed prints money.
   b. Investors buy stock.
   c. Imports and exports.
   d. The Treasury prints money.
   e. Productivity.
3. The US central banking system was created in:
   a. 1791
   b. 1816
   c. 1913
   d. All of the above
   e. None of the above
4. The FOMC determines:
   a. How much money to print
   b. Minimum wages
   c. Unemployment rates
   d. How many taxes to collect
   e. Which banks get money
5. The Fed's balance sheet expands when:
   a. A financial crisis occurs.
   b. The banks need more money.
   c. The Fed buys assets.
   d. The Fed sells assets.
   e. The Fed sells liabilities.
6. An economic paradox of enrichment occurs when:
   a. Too much money results in too little money.
   b. Wealthy people get richer and poor people get poorer.
   c. Wealth distribution is long-tailed.
   d. The economy is tipped in favor of bankers.
   e. The Fed's monetary policy is to expand.
7. The Fed's balance sheet as of 2013 was at a level considered:
   a. About right
   b. Historic
   c. Normal
   d. Only a small percentage of GDP
   e. The reason the economy was saved
8. Economic carrying capacity is an underlying financial capacity of:
   a. Homeowners
   b. Banks and savings and loans
   c. Economy's ability to sustain debt
   d. The Fed's balance sheet to expand
   e. The Fed's balance sheet to contract
9. The first electronic financial network in the United States was:
   a. FedWire using Morse code
   b. VISANet
   c. SWIFT
   d. TARGET
   e. CHIPS
10. The trend in financial networks is toward:
    a. Faster global networks
    b. Privatization
    c. Public–private partnerships
    d. Adoption of TCP/IP
    e. Less vulnerability
11. The *daylight overdraft* system uses a float called:
    a. The float
    b. M1 money
    c. Net debit cap
    d. Bank collateral
    e. Confidence in the bank
12. A credit crunch occurs when:
    a. Banks are afraid or unwilling to make loans.
    b. There is too little money to borrow.
    c. The central bank runs out of money to loan.
    d. The central bank is in a liquidity trap.
    e. The central bank withholds money from banks.
13. SWIFT is:
    a. Standards based
    b. International

c. A network infrastructure only
d. Member owned
e. All of the above

14. Which of the following applies to credit card networks like VISANet?
    a. They do not issue credit cards.
    b. They are financial clearinghouses.
    c. They handle the New York stock market exchange transactions.
    d. All of the above.
    e. None of the above.
15. The 3-D security protocol used by credit card networks is based on:
    a. PKI encryption
    b. Three layers or domains
    c. Internet protocols
    d. All of the above
    e. None of the above
16. The main advantage of a virtual currency is:
    a. Peer-to-peer networking
    b. Anonymity and low cost transactions
    c. Speculation
    d. Appreciation
    e. Consumer confidence
17. Virtual currencies like bitcoin are backed by:
    a. Gold
    b. Exchange rates
    c. Governments
    d. Merchants
    e. Confidence
18. Hackers are attracted to the global financial system because:
    a. It has become a vulnerable monoculture.
    b. International banking networks are poorly monitored.
    c. Cross-border transactions are cheap and easy to do.
    d. All of the above.
    e. None of the above.
19. Hot money is defined as:
    a. Contraband currency
    b. Counterfeit currency
    c. Speculative investment in search of higher returns
    d. A paradox of enrichment
    e. A liquidity trap
20. Liquidity trap occurs when:
    a. Expansionist monetary policy fails to stimulate.
    b. China buys US dollars.
    c. The euro/dollar exchange rate falters.
    d. People spend too much.
    e. Monetary policy devalues the dollar.

17.9 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. Ripple is another cryptocurrency that solves the double-spend problem differently than bitcoin. Explain ripple and compare it with bitcoin.
B. Compare EMH and FMH in terms of underlying assumptions about how the economy works. How are they similar? How are they different?
C. Compare and contrast ApplePay, PayPal, and the 3-D secure payment protocol.
D. Explain Figure 17.8 in your own words. Why is Figure 17.8b different than Figure 17.8a?
E. Explain Figure 16.3 of the previous chapter in terms of FMH.

REFERENCES

[1] Sinn, H.-W. and Wollmershäuser, T. Target Loans, Current Account Balances and Capital Flows: The ECB's Rescue Facility. *International Tax and Public Finance*, 19, 4, 2012, pp. 468–508.

[2] Lamport, L., Shostak, R., and M. Pease. The Byzantine Generals Problem, *ACM Transactions on Programming Languages and Systems,* 4, 3, 1982, pp. 382–401.

[3] Levchenko, K., Pitsillidis, A., Chachra, N., Enright, B., Felegyhazi, M., Grier, C., Halvorson, T., Kanich, C., Kreibich, C., Liu, H., McCoy, D., Weaver, N., Paxson, V., Voelker, G. M., and Savage, S. Click Trajectories: End-to-End Analysis of the Spam Value Chain. *IEEE Symposium on Security and Privacy*, Oakland, CA, 2011.

[4] McKinnon, R. and Liu, Z. Hot Money Flows, Commodity Price Cycles, and Financial Repression in the US and the People's Republic of China: The Consequences of Near Zero US Interest Rates. *Asian Development Bank Working Paper Series on Regional Economic Integration*, No. 107, January 2013.

[5] *The Economist*. The Dutch Disease. November 26, 1977, pp. 82–83.

[6] Corden W. M. Boom Sector and Dutch Disease Economics: Survey and Consolidation. *Oxford Economic Papers*, 36, 1984, pp. 362.

[7] Fama, E. Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, 25, 2, 1970, pp. 383–417.

[8] Peters, E. E. *Chaos and Order in the Capital Markets: A New View of Cycles, Prices, and Market Volatility*, 2nd ed, New York: John Wiley & Sons, Inc., 1996.

[9] Hurst, H. E. Long Term Storage Capacity of Reservoirs. *Transactions of the American Society of Civil Engineers*, 116, 1951, pp. 770–799.

[10] Sornette, D. and Johansen, A. Large Financial Crashes, *Physica A*, 245, 3–4, 1997, pp. 411–422.

[11] Lewis, T. G. *Book of Extremes: Why the 21st Century Isn't Like the 20th Century*, Cham: Springer, 2015.

18

STRATEGIES FOR A NETWORKED NATION

The foregoing chapters on complexity theory, network science, and the most essential CIKR sectors should equip the reader with knowledge and skills for making good homeland security policy. But knowledge and skills are inadequate without a strategy for securing the nation's CIKR at physical, cyber, and organizational levels. The following scenario-based approach emphasizes modern complex system's analysis in formulating a strategy going forward. In particular, it advocates a strategy based on increasing returns and network effects, shared responsibility, reducing friction, the concept of a national infrastructure corridor co-located with established rights-of-ways, and other techniques for reducing risk and increasing resilience.

Increasing returns and network effects remind us of Metcalf's law that says, "the power of a network increases with the square of number of nodes." This is another way of expressing synergy within a connected system that comes from connectivity itself. That is, capability exponentially increases simply by being connected. The capabilities of the whole are greater than the sum of capabilities of individuals. Increasing returns has been described as "the rich get richer"—the same self-organizing property of Gause's competitive exclusion principle. When applied to strategy, it means a good strategy leverages network effects to magnify resilience and the ability to tolerate stress.

Shared responsibility is a strategy of spreading risk across many parties and/or layers of a system. One of the best illustrations of this is the way the US Forest Services within the Department of the Interior spreads responsibility for fighting forest fires across local, tribal, state, and federal jurisdictions. When a major forest fire breaks out in California or Washington, every firefighting unit in the 11 Western states responds. The Price–Anderson Act that re-insures nuclear power plant owners against catastrophic failure of any one of the 104 nuclear power plants in the United States spreads risk across all owners and operators as well as the federal government. Shared responsibility means everyone in the network is responsible for everyone else. The weakest link becomes a liability for all.

Friction is found in most every complex system whether it be physical, cyber, or organizational. The modern world attempts to reduce friction using modern tools such as information technology (IT), lean management techniques, and other synergistic effects. Unfortunately, many governments have been slow to reduce friction in daily processes as mundane as registering to vote, obtaining driver's licenses and permits, and paying taxes. Homeland security strategies are at risk of increasing friction, rather than reducing it, as they pursue security. This is counterproductive. A homeland security strategy that increases friction is likely to fail. Instead, every new policy and regulation must be subjected to a test—does it increase or reduce friction? The answer must be that it reduces friction.

A common approach to reducing friction is the substitute IT automation in place of manual and face-to-face processes. The use of smartphone apps in place of taxicab hailing, making hotel reservations, making appointments, and health monitoring services are examples of friction lowering tools found everywhere online, but often lacking in governmental services. The lessons learned by e-commerce sites need to be

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

incorporated into various homeland security processes such that friction is reduced, not increased.

A number of other techniques of modern management and friction-free processes specific to each sector will be recommended in the following analyses. Specifically, the author recommends a national infrastructure corridor strategy that uses existing rights-of-ways to overcome NIMBY (Not In My Backyard) sentiment against infrastructure development. Existing highways and freeways are an obvious opportunity for co-locating, energy, power, and communications infrastructures. In this strategy, gas, electricity, electrical storage, and Internet packets travel along the same routes as highways—typically buried underground next to or under the roadway.

These strategies are recommendations for further discussion and thought. They are not intended to be complete. Rather, they are intended to be a starting point for further discussion. A brief outline and some details for the most critical infrastructure systems described in this book are given below.

18.1 WHOLE OF GOVERNMENT

The concept of whole of government was introduced in Chapter 1. It is a simple concept to grasp but often ignored by homeland security strategists. It begins at the top with the federal bureaucracy we know of as the Department of Homeland Security (DHS), which was created as an amalgam of agencies ranging in diversity from the Secret Service to FEMA. These agencies often overlap, compete for funds, and sometimes contradict other agencies. For example, responsibility and control of dangerous chemical cargo such as chlorine transported in railroad cars overlaps with the responsibility of the transportation sector, chemical sector, and various law enforcement agencies. The regulation of electrical power transmission lines across Native American sovereign nations clashes with federal, state, and local regulations. Funding at the local and municipal levels pits one county or city against adjacent counties and cities instead of promoting collaboration.

The role of DHS in state and local security has not been fully established as a solid strategy at the time this was written. A simple interpretation of DHS's role in securing local jurisdictions is that of a funding agency. (Other interpretations are that DHS is a law/regulation enforcement agency, a standards setting agency, a federal emergency management response agency, the czar of cybersecurity, and so on.) Assuming the primary role of DHS is to provide funding, how would the Department use funding to encourage increasing returns, shared responsibility, and friction-free operations?

The model proposed here is derived from the US Forest Service model of shared responsibility across a network of local firefighters. Two illustrative examples are given: encouraging preparedness against the threat of hurricanes in the Southeastern United States and response to a deadly pandemic within large metropolitan areas. Preparedness against hurricanes is a regional strategy, while response to deadly pandemics such as an outbreak of smallpox is a national strategy.

Hurricanes are almost certain to hit a region of the United States every year. It is not an uncertainty. Rather, it is a certainty without knowing exactly where and when the next hurricane will strike some part of the country. Unfortunately, individual counties and states are left to their own defenses. FEMA and private sector companies such as electric power utilities step in after a hurricane has occurred. There is no shared responsibility, nor is there any attempt to leverage capabilities across the country. Federal funding is perhaps the closest thing to shared responsibility, but other options exist.

Responses to hurricanes can employ increasing returns by organizing a network of hurricane responders located across the states most likely to be impacted by a hurricane. Like the US Department of Interior's US Forest Service, first responders from across the United States (or portions most affected) should respond to damages regardless of location. This also implies sharing of equipment, medical supplies, housing, and so forth.

Federal funding for preparedness would only be allocated to those members of the "hurricane network" to encourage collaboration across local and state boundaries. It may also be necessary to collect emergency response taxes from all utilities in the hurricane network region to be used by all impacted regions. This is similar to the Price–Anderson Act that makes each nuclear power plant share liability with the government for damages to any one of the plants.

The response to a pandemic such as an outbreak of smallpox in a densely populated city is another example of shared responsibility and how network effects may be used to support response. Required response time to an outbreak of smallpox is 3 days or less. It would take a cadre of 50,000 or more trained public health caretakers to respond in a timely manner to a smallpox attack in Manhattan, New York, where 8 million people might become infected. This capability does not exist in Manhattan or any other densely populated area of the United States.

A strategy based on network effects and shared responsibility would prepare 50,000 healthcare professionals throughout the United States with supplies and capability to respond quickly to any outbreak in any community, regardless of city, county, or state boundaries. Such a network must have access to rapid response capability including transportation to and from an infected region. Response would come from a national level effort rather than local jurisdictions.

18.2 RISK AND RESILIENCE

Risk and resilience are subjected to *prospect theory* that says humans are incapable of accurately judging risk. A majority of humans misjudge risk because of fear or irrational reasoning. For example, the odds of dying in an automobile accident are much greater than the chances of death by terrorist or airplane crash. And yet, most people fear terrorism and airplanes more than automobiles. Prospect theory explains why.

When subjects in a controlled experiment were asked to choose between accepting $10 outright and taking a 10% chance of winning $100, more than 80% elected to accept $10, even though the risk is identical. This is known as risk avoidance, because subjects accepted the certainty of $10 rather than the chance of winning $100.

When the same subjects were asked to pay $10 or accept a 10% chance of paying $100, most subjects accepted the risk and elected to pay $100 ten percent of the time. This is known as risk-seeking behavior because subjects accepted a potential loss of $100 rather than a certain loss of $10.

In both cases, risk is the same—$10, because risk is expected gain or loss. Using 10% as the probability in the formula for risk, we get (0.1)($100) = $10, which is the same as accepting $10 or potentially losing $100 ten percent of the time. In a rational world, there is no difference between the two scenarios. But in the prospect theory world, there is a significant difference because of human irrationality.

Prospect theory explains why people are unable to judge risk. It depends on an individual's situation and perception of control over the situation. A wealthy person is more risk seeking than a poor person, because loss is less damaging to a wealthy person. Wealthy people are less likely to buy insurance, because they can afford to replace a damaged asset. A poor person, on the other hand, is wise to buy insurance as a hedge against disaster.

Additionally, control over one's situation tends to give gamblers more confidence in the outcome. A person confident in his or her driving skills is more likely to risk driving an automobile in heavy traffic than riding in an airplane. The perception of control misguides the confident driver into thinking travel by automobile is less risky than travel by an airplane that is not under the driver's control. On the other hand, a person that is afraid to drive an automobile in heavy traffic is more likely to take a bus or public transportation for fear of an accident.

Prospect theory has a major impact on homeland security, because it generally drives citizen taxpayers toward risk avoidance on the upside and risk seeking on the downside. That is, the general public is reluctant to spend personal money on flood insurance but favor spending taxpayer money to protect commercial air travel from terrorism. This is largely due to the perception that we have control over hurricane preparedness but not terrorism. A rational person should be more concerned with flooding than terrorism. And yet, the opposite is true because of prospect theory. As a result, the federal government subsidizes flood insurance and spends vast sums of money on counterterrorism and TSA.

From a policy point of view, prospect theory makes it very difficult to convince people to pay for preparedness. It is easier to convince people to pay for a catastrophe after it happens than beforehand. Furthermore, it is difficult to convince people to buy insurance against a low-probability, high-impact catastrophe. Most people do not carry insurance against death by asteroid. Similarly, large corporations are more likely to transfer risk to an underwriter than spend profits on prevention. It is not unusual for an electrical power utility to delay maintenance as a prevention measure and buy insurance against outages, instead.

Given that most catastrophic events are governed by long-tailed exceedence probability that places black swan events at the far right hand side of the risk profile, the optimal strategy for resilience is to invest heavily in *responding* to low-consequence, high-frequency events such as fires and hurricanes and invest in *preventing* high-consequence, low-frequency events such as nuclear accidents and asteroid impacts. Homeland security strategy should minimize impact of low-consequence incidents and maximize prevention of high-consequence incidents. For example, prevention of an existential event such as asteroid collisions means development of rockets to destroy asteroids. Response to nonexistential events such as floods, fires, and hurricanes means providing EMS teams with equipment and training to effectively respond to "natural hazards."

This dual strategy of prevention and response should be a balanced strategy. High-frequency events like floods and fires should still be prevented as much as possible by building levies and legislating strict building codes. Response to existential events such as asteroid collisions and nuclear accidents is still needed, but these capabilities evolve over long periods of time.

Regardless of the prevention versus response strategy suggested by long-tailed exceedence distributions, homeland security risk assessment must be quantified in order to avoid the mistakes caused by prospect theory. Risk is found where it can be quantified and not necessarily where fear and uncertainty suggest it is. Emotional decisions based on fear and uncertainty must be set aside. Only rational quantifiable risk assessment can properly inform risk-informed decision-making.

18.3 COMPLEX AND EMERGENT CIKR

Most of the CIKR examples studied in this book illustrate the effects of self-organization. Over time, most CIKR self-organize and become more structured due to a number of

factors such as cost, efficiency, and regulation. For example, the communications sector has become a hub-and-spoke network because of the 1996 Telecommunications Act. Most supply chains are highly structured to reduce costs and streamline operations. Self-organization increases risk of catastrophic cascading and reduced resilience. What if self-organization is reversed? That is, restructuring to reduce spectral radius also reduces risk and increases resilience.

Figure 18.1 illustrates the impact of rewiring the links of a self-organized network like the one in Figure 10.7. The PML risk and spectral radius of this network were originally 876 and 6.95, respectively. After rewiring as shown in Figure 18.1, PML risk and spectral radius are reduced to 624 and 6.19, respectively. Resilience increases from 4.6 to 5.56 per the definition of cascade resilience. The tail of the exceedence probability distribution is shortened, and fractal dimension increases from 0.96 to 1.58. This "pushes" PML risk to the left or closer to zero.

Restructuring by rewiring the links of a network is one method of reducing self-organization, but it is not the only method. Figure 18.1 was obtained by switching link assignments such that cascade PML risk is reduced. A similar process might reduce flow risk, spectral radius, and so on.

18.4 COMMUNICATIONS AND THE INTERNET

The hub-and-spoke structure of both communications and Internet systems suggests that they are extremely self-organized. Communication hubs are located in a handful of carrier hotels, and the spectral radius of the Internet is very large, perhaps greater than 150. This means that malware travels fast and far throughout these super-spreading networks. The implication is that the hubs should be dissipated and the topology of the communication infrastructure "randomized" such that its structure more closely resembles the interstate highway network with its very low spectral radius.

Rewiring the global Internet is unlikely to happen, however. The alternative is to organize the communications and Internet CIKR along the lines of the nuclear power industry with its shared responsibility. Combating malware is a whole of community challenge requiring cooperation among stakeholders—both government and private sector owners and operators. Full participation from both public and private sectors is required because of the entanglement of both.

A shared responsibility model might be structured along the lines of the US National Forest firefighting agencies, whereby all participants play a role in response. More likely,

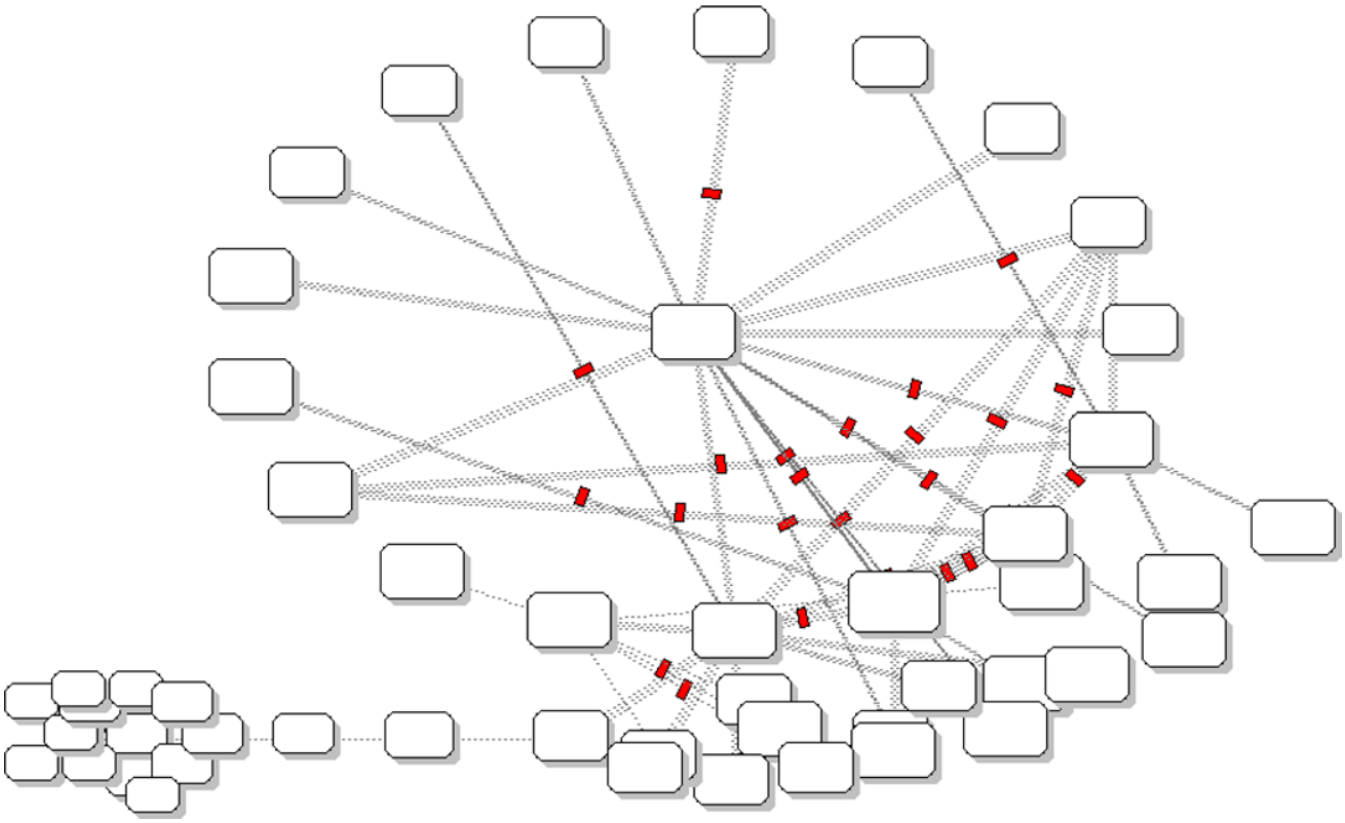


FIGURE 18.1 Rewiring the network of Figure 10.7 to reduce PML risk improves both risk and resilience. The marked links have been reassigned nodes to reduce the centrality of the hub node, which reduces spectral radius. In turn, lower spectral radius means lower PML risk.

the organizational structure works best when the strategy is to prevent the spread of malware from server to server, agency to agency, and e-commerce to e-commerce site.

An alert system built into BGP (Border Gateway Protocol) could carry malware alerts and suspicious activity reports along with data. Recall that the BGP system, like the Interstate Highway System, connects major Internet hubs to one another and carries the bulk of data over long distances. It is a natural conduit for both malware and countermeasures.

Additionally, the major waypoints along the BGP routes should be carefully protected as well as the major DNS servers like 1.1.1.1. These are the most likely hubs to receive and forward malware. Therefore, they should be subjected to extreme oversight and scrutiny. At some point, these critical assets may be regulated and required to implement a standard set of countermeasures.

18.5 INFORMATION TECHNOLOGY (IT)

The question for this sector is whether or not to regulate it. Up until approximately 2017, government adopted a laissez-faire attitude regarding information technology. But when China began to challenge the United States in high-tech areas like AI and 5G communications, the laissez-faire approach began to turn into bewilderment and concern, but without much focus or direction.

Regardless of the weak leadership of Western nations regarding IT, homeland security must be gravely concerned about safety and security in high-tech devices and products. For example, the state of California requires original equipment manufacturers to pre-install unique passwords in IoT devices. Adoption of a shared responsibility across IT service providers such as cloud services might follow the model of the Safe Drinking Water Act. This act segments drinking water into two classes—small and large. Water sources used by fewer than 3300 people are considered small and subject to different public health regulations than large systems. Large suppliers are subjected to much more demanding requirements in terms of health safety and terrorism-related risks.

Similarly, large IT systems may be subjected to much more demanding standards than small IT systems. Large-scale cloud computing platforms may be required to install extensive countermeasures more than small operators. Furthermore, large-scale platform owners and operators may be subjected to fines for noncompliance. For example, the European General Data Protection Regulation (GDPR) allows the EU to fine e-commerce owners and operators up to 4% of annual revenues for not complying with the GDPR.

What might these requirements be? The GDPR protects user's privacy much like the Health Insurance Portability and Accountability Act of 1996 (HIPAA) protect health-related information on consumers. But it does not go far enough in terms of surveillance capitalism and deeper privacy issues. For example, GDPR requires notification of cookies and consumer information collection, but it does not forbid cookies and data collection. It does not impose fines for successful exfiltration of personal information from e-commerce servers when attacked by criminals. And it does not require encryption of communications between and among consumers.

Both US laws and EU regulations on personal information collection do not address the root causes of loss of information—inadequate countermeasures and procedures. For example, there is no requirement to use FIDO2/WebAuth to authenticate users, minimize user's account information held in online databases, or encrypt password files, and so on. Simple mechanisms long known to minimize risk are not required, which means that some owners and operators do not implement even the simplest countermeasures.

The debate over strong encryption may not be over, but it is clear that backdoors and methods of circumventing strong encryption result in weak or no encryption. That is, strong encryption is a binary choice—you have it or you do not. There are obvious arguments in favor of law enforcement having the ability to bypass strong encryption. But this exposes the larger community of users to risks that far outweigh the benefits to law enforcement.

Strong encryption is a basic right of consumers. However, there are downsides to strong encryption beyond the demands of law enforcement. Strong encryption allows terrorists and criminals to hide their activities. It also allows filter bubbles to form, fake news to proliferate, and fosters antisocial activity without exposure to online tools used by social networks to stop the spread of misinformation. Users of WhatsApp, Instagram, and blue iMessage can hide their activities under the umbrella of strong encryption, which facilitates the spread of misinformation.

18.6 SURVEILLANCE CAPITALISM

In line with regulated IT systems is the problem of surveillance capitalism—profiting from collecting and selling highly targeted personal information on consumers. Social networks have exploited highly detailed personal information gleaned from their e-commerce sites for decades prior to the *Cambridge Analytica* scandal with Facebook.com. As described in Chapter 9, the solution is regulation, but what kinds of data and behaviors should be regulated? Regulation is difficult because of two major factors:

- Regulators do not always understand technology and its subtlety.
- Conflicts with freedom of expression.

Technology is very powerful and subtle in how it exfiltrates private information from users. For example, much can be

learned about a person's private life by tracking locations, observing how a consumer uses his or her smartphone, and various forms of aggregation. An insurance company may deny coverage if it knows a person consumes a certain drug.

Regulation is not a complete answer, but it is a starting point. For example, social networks and e-commerce sites such as Amazon.com may be required to separate surveillance data from consumer data through a third-party mechanism. Suppose company X is allowed to collect encrypted surveillance data, anonymize and aggregate it, and then pass it on to social network or e-commerce site for the purpose of making recommendations and giving feedback. The process is reversed—recommendations and feedback return to the consumer via company X. In this way the social network and e-commerce company only have profiles representing demographic and psychographic data and not names and addresses.

Furthermore, company X cannot sell surveillance data without anonymization and aggregation, following Institutional Research Board (IRB)-like rules governing human research. This includes exempting children, not targeting minorities, and not extracting information obtained by implication and deep learning algorithms, and so on.

18.7 INDUSTRIAL CONTROL SYSTEMS

The primary issue with industrial control systems, SCADA, and energy management systems (EMS) is age, extremely long replacement cycles, and overly relaxed controls. Remote terminal units (RTUs) are often installed without changing factory set passwords that are simple to begin with. Furthermore, SCADA systems are frequently running on older Microsoft Windows operating systems that are not patched or no longer supported by Microsoft.

Any strategy attempting to harden industrial control systems must address the inadequacy of old devices and old operating systems. These must be replaced by newer, hardened RTUs and operating systems. Replacement is a vast problem because of the millions of obsolete devices and the high cost to replace major portions of these systems. It is unlikely owners and operators will do so without financial incentives or regulation. Replacement will be expensive and take time.

Meanwhile, these old and unsecured systems must be quarantined by air gaps and special controls. For example, most network management systems use the SNMP (Simple Network Management Protocol) to check on operations. Perhaps a special ruggedized SNMP management system is needed expressly for SCADA.

18.8 ENERGY AND POWER

Energy and electric power systems are extremely important special-purpose industrial controls. In addition, they are confronted with an uncertain future due to climate change and the ensuing transition from fossil fuels to solar and wind power. Essentially, the energy/power sectors are undergoing a complete revision, but the direction of this revision is currently uncertain. Is it headed for distributed renewables, or is it going to end up being a centralized mix of renewables and existing sources of energy?

Two scenarios are possible and should be considered. Distributed renewables typically means local microgrids operate from local energy sources such rooftop solar, local windmill power, and batteries. Net metering permits homeowners and small-scale businesses such as shopping malls and factories to produce their own energy. Surplus energy is stored locally in batteries. Even more surplus is fed back into the microgrid for others to use.

Over time, distributed generation via renewables reduces reliance on large regional grids. Only local distribution is needed. Peaks and valleys in demand are offset by storage. The vehicle-to-grid (V2G) vision integrates electric vehicles with their large storage capacities into the microgrid model so that electric automobiles become part of balancing supply and demand—Area Control Error (ACE). The V2G model has been shown to work on a small scale. This model reduces the importance of large energy suppliers and regional utilities. It depends more heavily on consumers adopting renewables on their own. Is it the future?

The mixed renewables with centralized generation scenario is currently a competitor to the fully distributed generation model. In this model, large regional and perhaps even national grids balance energy supply and demand via large solar and wind generation "plants." Large utility-owned and operated storage and traditional gas peaker plants simplify balancing across major regions of the country. Large energy suppliers and utilities own and operate highly centralized and massive solar and wind farms. Very little demand is filled from rooftops and consumer storage batteries. Transmission and distribution through wires becomes even more important to maintain ACE near zero.

For example, in 2019 Missouri regulators approved the *Grain Belt Express* transmission line—a 780-mile overhead direct current (DC) transmission line designed to bring 4000 MW of wind power from Kansas through Missouri, Illinois, and Indiana, and then farther east, into the Eastern grid via DC. A similar project is proposed from Iowa east through Illinois involving both wind and solar.

Regardless of which scenario wins, both models must cope with cybersecurity issues as described in this book. This will require much stricter regulation than currently enforced by government. Energy and power utilities are likely to be required to conform to the NIST-CSF and other frameworks going forward. Voluntary compliance is unlikely to be adequate.

One visionary approach to energy and power CIKR is to co-locate (renewable) generation, storage, and transmission along energy corridors. These corridors may actually already

exist in the form of the Interstate Highway System and railways. For example, electric power transmission, gas and oil pipelines, Internet fiber, and batteries for storing electric power can be co-located along state-owned rights-of-way (highways). Knitting together a power grid containing 1,000 batteries approximately every 40 miles along the 40,000 miles of the Interstate Highway System would prove to be an extremely resilient power grid. Such a grid theoretically could collect electric power from wind and solar generators located in states with high wind and solar potential and deliver it anywhere in the country. And, because the Interstate Highway System is extremely resilient with many alternate routes, the network would be extremely resilient, too.

18.9 GLOBAL PANDEMICS

Public health is inadequately staffed and funded in most countries, and especially in the United States, where there is no universal healthcare, hospitals are optimized for profit, and vaccination requirements are lax or nonexistent. Many communities are without full-time EMS or are funded by local charities. An outbreak of a serious disease in a major metropolitan area would seriously tax response.

The Roemer model and HSPD-21 is an adequate strategy, if it is carried out at state and local levels. Recall from Chapter 14 the strategy has four pillars: national bio-surveillance, stockpiling and distribution, mass casualty care, and community resilience. Unfortunately, most of the United States falls short of implementing this strategy. Prospect theory may be the reason.

A modern risk sharing approach to each of the pillars is proposed, here. The idea is simple—create a network of systems that collaborate to achieve network effects. First, bio-surveillance through reports from hospitals, clinics, and doctor's offices to the CDC should be enhanced by predictive analytics obtained from social media companies and wearable computer vendors. Google.com, for example, has accurately predicted the outbreak of Asian flu by mining data obtained from consumer searches for cold remedies. The spread of deadly viruses such as Ebola can be predicted by mining cell phone locations. The Apple Watch collects fitness data on its users. These modern methods of prediction and tracking should be formalized and implemented by the CDC.

Stockpiling and distribution is a major problem compounded by the fact that some drugs are manufactured by a single company that is at risk of ceasing operation during long periods between demand. Demand is extremely bursty, even for common remedies such as cipro and penicillin. There is no solution to this problem short of subsidies paid to suppliers to stay in business. Someone must pay for readiness even when readiness is not used.

Distribution is a major problem for most densely populated communities where either hoarding of drugs or inadequate access is problematic. A mass exodus from large cities such as Los Angeles following announcement of a serious contagion such as smallpox would block road networks, keeping emergency vehicles from delivering food, water, and drugs. Some communities are partnering with retailers to preposition drugs close to residential areas, but this strategy may be easily overwhelmed.

Because of the issues described above, it is essential that communities develop emergency food, water, and medicine supply and distribution plans far in advance of a pandemic. This "whole of community" strategy must incorporate solutions to the problems of surveillance, stockpiling, distribution, and mass casualty care. Public schools, government buildings, and sports arenas must be made capable of handling mass casualties, rationed supplies, and care.

18.10 TRANSPORTATION AND SUPPLY CHAINS

Road and rail transportation systems are the most robust and resilient CIKR of all sectors studied in this book. Multiple alternate paths exist between almost any two points on the map. If one route is destroyed or congested, a secondary route exists, even though it may be longer or slower. Unfortunately, this is not the case for most supply chains.

For very reasonable economic reasons, supply chains are highly optimized. As a consequence, they typically lack multiple sources, multiple routes between producers and consumers, and just-in-time inventory, which means there is minimal surge capacity. Furthermore, supply chain managers are not motivated to increase redundancy or provide surge capacity because of cost. Supply chains are perhaps the most fragile of all CIKR.

Supply chain managers typically do not include the cost of asset failure in their optimized systems. That is, they assume nothing serious will occur. A lost hour or two can be made up with overtime. The impact on production from a delayed shipment incurs minor costs. Unfortunately, this is not the case in practice. For example, the Fukushima East Asia earthquake disaster heavily and seriously impacted automobile manufacturing all over the world. It damaged and closed key ports and some airports. It disrupted 20% of the world's semiconductor products that go into global products like the Apple iPad and wings, landing gears, and other major parts of Boeing's 787 Dreamliner. Automakers Toyota, Nissan, Honda, Mitsubishi, and Suzuki temporarily suspended production. A total of 22 plants in the area, including Sony, were shut.

But the disaster also suggested a precaution against future supply chain disasters. Since the 2011 disaster, global auto suppliers changed the way they produce and source the 30,000 parts required to assemble a single car, by raising stocks, diversifying production, and creating alternative manufacturing capabilities.

Supply chain owners and operators have a choice—accept long-tailed risk or suboptimize the chain making it more redundant, increase surge capacity, and use multiple suppliers. Either way, costs will rise.

18.11 BANKING AND FINANCE

Two major threats face banking and finance—poor fiscal management by governments that lead to collapse like the 2008–2009 financial collapse due to the paradox of enrichment and systemic cyber attacks that empty out banks. The fiscal policies that led to the 2008–2009 financial collapse cost the US upwards of $12.8 trillion. This loss far exceeds the cost of any blackout, hurricane, or earthquake that has occurred over the past 100 years.

"Estimated actual gross domestic product (GDP) loss from 2008 to 2018, of $7.6 trillion. This is the cumulative difference between potential GDP—what GDP would have been but for the financial and economic crises—and actual and forecast GDP during the period. Estimated avoided GDP loss from 2008 to 2012 of $5.2 trillion. This figure is the estimated additional amount of GDP loss that was prevented only by extraordinary fiscal and monetary policy actions."[1]

The economic loss due to bank robbery by cybercriminals is less clear. One report claims in 2018 that the annual cost of all cybercrimes was $1 trillion.[2] While this is much less than the financial crisis cost, it is a recurring cost. Additionally, it is three times greater than the $300 billion lost due to natural disasters.

Most governments have no strategy for coping with these losses. Cybersecurity is left mainly to the private sector. Fortunately, the private sector has been very responsive to prevention of banking crimes. Authentication has significantly improved due to FIDO2 and WebAuth, token-based credit cards like the Apple iPhone mechanism, and the general trend toward authentication without passwords.

Most financial transactions across borders and through inter-banking systems such as SWIFT use VPN security. But some governments require VPN operators to collect and store activity information on their users. This negates the virtue of VPNs, especially if personal information is stored by the VPN. Authoritarian governments can demand and get user's activity information.

[1]https://finance.yahoo.com/blogs/daily-ticker/2008-financial-crisis-cost-americans-12-8-trillion-145432501.html

[2]https://www.globalsign.com/en/blog/cyber-bank-robberies-contribute-to-1-trillion-in-cybercrime-losses/

The EU GDPR prohibits collection of activity information by VPNs if users opt out. If the VPN conforms to the GDPR, consumers may opt out of the collection of personal information such as where they are located, where they have traveled, what they have bought, and where they bought it. Conformity to the GDPR is the current best strategy for protecting the banking system.

Anas Baig writes, "In my opinion, GDPR is a great step that's been taken towards making the Internet a safe place for humans from all walks of life. It helps ensure ultimate privacy and security for personal information of all users, irrespective of their usage of the Internet. Not only that, it also regulates online utilities and services, and controls the amount of data they can use and share."[3]

18.12 DISCUSSIONS

The following questions can be answered in 500 words or less, in slide presentation, or online video formats.

A. Real options analysis is a type of scenario planning whereby options are analyzed and compared to determine a best strategy going forward. Perform an options analysis on the energy and power sectors in terms of the choice between centralized and distributed generation. What policies apply in each case?
B. A number of frameworks exist specifically for each CIKR sector. They are generally checklists of proper things to consider and analyze. How would you compare and contrast risk-informed decision-making with typical frameworks such as the NIST-CSF? How do they impact strategy?
C. Cybersecurity was largely left up to the private sector until 2016–2017 when surveillance capitalism and meddling in US elections were revealed. What strategies and policies should the US government enact to protect consumers and voters in the future?
D. This chapter emphasizes the use of network effects and shared responsibilities as fundamental strategies. What other fundamental strategies might be employed to protect CIKR as well as respond to CIKR incidents?
E. Global climate change may be the biggest threat to CIKR over the next 50 years. What strategy should governments use to combat the effects of climate change on CIKR?

[3]https://www.globalsign.com/en/blog/what-gdpr-means-for-vpn-providers-and-users/

APPENDIX A

MATH: PROBABILITY PRIMER

The publication of Geronimo Cardano's (1501–1576) *Ars Magna* in 1545 is generally recognized as the beginning of modern mathematics even though his most important works were not published until after his death [1]. Cardano's mathematics was somewhat at odds with Isaac Newton's (1642–1727) deterministic universe in which an apple falls from its tree in exactly the same way every time it falls. Its path is predictable in every detail.

Newton's deterministic model cast a shadow over Cardano's nondeterministic model for over a century. It was perfect for explaining some phenomena but completely inadequate to describe many other everyday phenomena. For example, Newtonian mathematics failed to explain accidents, diseases, and games of chance. Newton's model could not predict *when* the apple would fall. On the other hand, Cardano turned uncertainty and games of chance into a practical application of mathematics. According to Wikipedia, "Cardano was notoriously short of money and kept himself solvent by being an accomplished gambler and chess player. His book about games of chance, Liber de ludo aleae ('Book on Games of Chance'), written around 1564 but not published until 1663—[during Newton's lifetime]—contains the first systematic treatment of probability, as well as a section on effective cheating methods."[1]

Cordano's ideas were developed even further by a young genius named Blaise Pascal (1623–1662). Pascal built and sold 50 mechanical calculators by the age of 18. In 1654 at the age of 30, Pascal vastly improved on a solution to a problem posed by the famous mathematician Pierre de Fermat and relayed to him by his mentor Chevalier de Méré—the *problem of the points*. And in the process, Pascal invented *probability theory*, which laid the foundation for *risk analysis*. More profoundly, Pascal's invention formalized Cardano's new idea—that the world is not entirely deterministic, nor is it entirely beyond the understanding of humans. Pascal's breakthrough came by an indirect means—he invented probability theory to answer Fermat's question of how to win games of chance. Pascal turned the study of chance into a mathematical science and created the foundation of nondeterministic sciences.

[1]https://en.wikipedia.org/wiki/Gerolamo_Cardano

A.1 A PRIORI PROBABILITY

The *problem of the points* asks, "What is the likelihood of winning a game of chance given that we know the partial-play scores of two players?" For example, in tennis the first player to exceed 40 wins. But if the score is 40–30, what is the probability of coming from behind and winning? This game has more than one possible win–loss outcome, depending on skill and luck. Unlike Newton's falling apple, the win–loss outcome of this game can differ each time it is played.

Pascal's solution is based on a simple assumption that *probability* is the number of ways of *winning* divided by the total number of ways of *winning and losing*. To come from behind and make two points before the leader with 40 points scores one more point, our underdog has to be either very

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

skilled or very lucky. Assuming equal skill, the chance of scoring two points before the leader scores one point is identical to the chance of getting two heads in a row when tossing a coin twice. Let H and T represent heads and tails, respectively. Then according to Pascal's newfound science, there can only be one HH outcome from two tosses: HH, HT, TH, or TT. Therefore, the probability of winning the tennis match when the underdog is down by 40–30 is 0.25, because only one of the four possible ways of finishing the game ends with the underdog reaching the winning score, first. In general, if #HH, #HT, #TH, and #TT are the number of HH, HT, TH, and TT runs, respectively, then the probability of HH, Pr(HH), is ¼:

$$\Pr(\mathrm{HH}) = \frac{\#\mathrm{HH's}}{\#\mathrm{HH} + \#\mathrm{HT} + \#\mathrm{TH} + \#\mathrm{TT}} = \frac{1}{4}$$

According to Pascal, when players are evenly matched, the difference between winning and losing is pure luck. More importantly, he showed that pure luck could be represented precisely by mathematics. Pascal defined *probability* as a number lying in the interval zero to one [0,1]. Zero represents an *impossible outcome*; one represents a *certain outcome*; and a fraction in between represents the likelihood of the desired outcome. Probabilities are sometimes written as a percentage. For example, 0.5 may be written as 50% and 0.05 as 5%.

Pascal's probability also represents *uncertainty* when it falls between zero and one. For example, 0.5 represents maximum uncertainty, because the desired outcome is no more likely to occur than not occur. Tossing a balanced coin is equally likely to turn up heads (H), as it is tails (T). Chances are 50–50% that one coin will land on heads or tails, 25–75% that two heads (HH) will happen in two tosses, 12.5–87.5% that three heads (HHH) will happen in three tosses, and so on.

The *odds* of one head (H) occurring in one toss is 1 to 1, two heads (HH) in two tosses is 1 to 3, and three heads (HHH) in three tosses is 1 to 7, because probability and odds are related by the ratio p to $(1-p)$. Therefore, when p is ½, $(1-p)$ is also ½, so ½ to ½ is 1 to 1; when $p = ¼$, $(1-p) = ¾$, so ¼ to ¾ is 1 to 3; and when $p = 1/8$, $(1-p) = 7/8$, so 1/8 to 7/8 is 1 to 7 odds. In general, if an event happens with probability p, it does not happen with probability $(1-p)$. So odds, o, is the ratio of an event happening to not happening:

$$o = \frac{p}{(1-p)}$$

Pascal generalized the problem of the points so it could be applied to all possible cases when the laggard is one, two, three, …, n points behind the leader. Here is how it worked. Consider Table A.1—a list of all possible combinations of heads (H) and tails (T) in four coin tosses. A head is a point for the laggard, while a tail is a point for the other player. Pascal assumed the coins were fair, so each toss of a coin produced H half of the time. Thus, all 16 of the combinations are equally likely. However, the number of heads (H) occurring in each combination of four tosses differs, as shown in the enumerated table. For example, 0 H's occur in 1 combination; 1 H occurs in 4 combinations; 2 H's occur in 6 combinations; and so forth. Thus, the probability of 0 H's is 1/16, 1 H is 4/16, 2 H's is 6/16, and so on. Table A.1 data is plotted in Figure A.1 as the probability of H's occurring in four tosses. This is known as the *probability distribution* or *density function* for tossing four coins. It is a graphic version of the information in Table A.1.

Pascal's method of solution became known as *Pascal's triangle*, because of the way patterns of heads and tails mount up in combinations. Table A.2 shows Pascal's triangle for all possible combinations of tossing four coins, each with an equal likelihood of landing heads up. The last row (#4) contains the numbers: 1-4-6-4-1, which equals the number of combinations with 0-1-2-3-4 H's in Table A.1. Dividing by the number of combinations, 16, yields the probability a combination containing 0, 1, 2, 3, or 4 heads. This is also the probability of coming from behind by 0, 1, 2, 3, or 4 to win when lagging the leader. And finally, the probability of coming from behind to win is also given by the distribution in Figure A.1. This distribution is known of as the *binomial distribution*, because each coin toss has two possible outcomes—H or T.

Cardano and Pascal invented the modern theory of probability by considering the ratio of one particular event to all possible events. They used combinatorial mathematics to count all possible "favorable outcomes"—the desired combination. Thus, Pascal defined probability as a simple ratio:

$$\Pr(x) = \frac{\mathrm{Number}(x)}{\mathrm{All}}$$

where Number(x) is the number of combinations containing x and All is the total number of combinations. For example, in Table A.1 the probability of 1 head occurring in 4 tosses Pr(1) is 4/16. This shows up as 0.25 in the distribution of Figure A.1. Similarly, the entire distribution of Figure A.1 is obtained by enumerating all probabilities:

$$\Pr(0) = 1/16 = 0.0625$$
$$\Pr(1) = 4/16 = 0.2500$$
$$\Pr(2) = 6/16 = 0.3750$$
$$\Pr(3) = 4/16 = 0.2500$$
$$\Pr(4) = 1/16 = 0.0625$$

TABLE A.1 There are 16 possible combinations of H and T in four coin tosses

Combinations	#H's	#Occurrences	Probability
TTTT	0	1	1/16
HTTT			
THTT	1	4	4/16
TTHT			
TTTH			
HHTT			
HTHT			
HTTH	2	6	6/16
THHT			
THTH			
TTHH			
THHH			
HTHH	3	4	4/16
HHTH			
HHHT			
HHHH	4	1	1/16

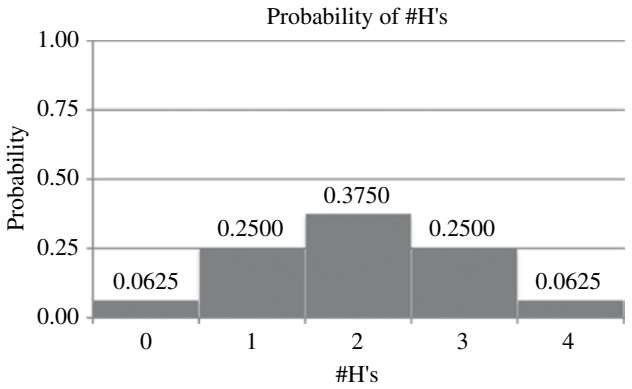


FIGURE A.1 Probability distribution for the number of heads occurring in four coin tosses.

Pascal's method of calculating probability is considered an a priori approach because it is predictive—it calculates the probability of a possible event, even before it happens. Various a priori methods of prediction continue to be used today. For example, we can calculate the likelihood of a terrorist attack before it happens by considering all the ways an attack can take place, but the answer depends on how you frame the question. For example, Michael Rothschild, a former professor at the University of Wisconsin, calculated that if terrorists entirely destroyed one of America's 40,000 shopping malls each week, the odds of being there at the wrong time would be about one in a million. If terrorists hijacked and crashed one of America's 18,000 commercial flights each week, the odds of being on the crashed plane would be 1 in 135,000.

If a 9/11-sized attack occurred every year, your 1-year odds of being part of the attack would be 1 in 100,000, and your lifetime odds would be about 1 in 1,300 (300,000,000 ÷ 3,000 = 100,000 ÷ 78 years = 1,282). Similarly, the odds of dying in a car accident is 1 in 6666, so that the likelihood of dying in a 9/11-sized terrorist attack is much lower than your odds of dying in a car accident, by walking across the street, by drowning, in a fire, by falling, or by being murdered [2].

TABLE A.2 Pascal's triangle yields the probability of 0, 1, 2, 3, 4, … H's in four tosses

#H's	Pascal's triangle				
0	1				
1	1	1			
2	1	2	1		
3	1	3	3	1	
4	1	4	6	4	1

A.2 A POSTERIORI PROBABILITY

Pascal's a priori method of calculating probability is based on combinatorial enumeration but ignores historical facts. For example, it ignores how many times in the past a tennis player has won when leading by 40–30. It also ignores the history of terrorism, earthquakes, floods, and cyber attacks. A priori means *before it happens*, which in turn means predicting the future by enumerating all possibilities. However, as shown above, enumerating all possibilities can be problematic.

Another 100 years would pass before a priori probability theory was extended by a posteriori probability theory—predictions based on past events and historical evidence. French mathematician and astronomer Pierre-Simon, marquis de Laplace (1749–1827) asked, "What is the probability the sun will rise tomorrow?" Laplace conjectured that future events are a consequence of past events and therefore their likelihood of occurring in the future can be calculated by simply counting the number of similar events that have happened in the past and dividing by all possible opportunities for the events to have happen over the same period of time.

Here is Laplace's result. Let S be the number of times in the past that the sun appeared on schedule. Then the probability that it will rise again is (S+1)/(S+2), according to Laplace. Given that the sun has never failed to rise, this number is very close to 1.0 or certainty. Conversely, the probability that the sun will *not* rise tomorrow is 1.0—(S+1)/(S+2) or very close to zero. But it is not exactly zero. Why?

Laplace's sunrise problem illustrates an important development in thinking about chance. His notion of chance events is based on evidence, while Pascal's combinatorial enumeration is based on mathematics. Laplace's interpretation contains an element of uncertainty, while Pascal's does not—an important and significant difference.

The presence of uncertainty in Laplace's model explains why the probability of the sun rising is not exactly 100% and the probability of it *not* rising is not exactly zero. This "side effect of uncertainty" is called Laplace's *rule of succession*. It says that when nothing is known about past performance, the probability of any event occurring in the future is (0+1)/(0+2) = 50%. In other words, the event either happens or not, with equal probability. As pointed out above, 50% represents "maximum ignorance" in situations where nothing is known about the likelihood of an event. As evidence mounts to the contrary (the sun has risen billions of times), the residue of uncertainty diminishes, but a small amount always remains in the estimate. Laplace's probability of the sun rising tomorrow is 99.9999…%, leaving a small lingering fraction of uncertainty.

Laplace's method has been criticized because of that famous stockbroker caveat, "past performance is no guarantee of future performance." And indeed, this is a valid criticism. Even after a stock has risen 1000 days in the past, there remains a residue of uncertainty that it will rise once again tomorrow. Laplace might respond to this criticism by arguing that accuracy can be (partially) improved by simply collecting more evidence!

Laplace's method is readily applicable to the problem of estimating the probability of a successful terrorist attack, given T attempts and S successful attacks in the past. Pr(successful terrorist attack) = (S+1)/(T+2). For example, between 2001 and 2011, there were S = 3 successful attacks in the United States, out of 26 attempts. Thus, the probability of a successful (future) attack, based on the past, is (3+1)/(26+2), or 14%. What happens to this estimate when more successful or unsuccessful attempts are recorded? Laplace would say that the estimates simply get better as the number of observations increase. That is, a posteriori probability estimates contain both truth and false positives and negatives. Uncertainty is reduced to zero only after an infinite number of data points have been observed and recorded.

In contrast to a priori probability, a posteriori probability uses historical data to calculate the probability of future events. A priori probability distributions are precise mathematical functions, such as the binomial distribution, while a posteriori probability distributions are empirically derived histograms. Recall that a histogram is a plot of frequency versus events, like the one shown in Figure A.2.

A.3 RANDOM NETWORKS

A priori and a posteriori probability estimates are used throughout this book to understand how infrastructure systems form and evolve over time. The fundamental tool is *network science*—the study of applied graph theory. Networks are collections of nodes and links representing a water supply system, power grid, Internet, or transportation system. Nodes can be anything—buildings, bridges, computers, airports, pumping stations, transformers, and so on. Links can also be anything as long as they connect nodes—roads, wires, rivers, pipes, routes, relationships between pairs of people, and so on.

Analysis of a CIKR system begins with a *network model* of the system. The nodes and links in an infrastructure system such as a power grid containing power generators,

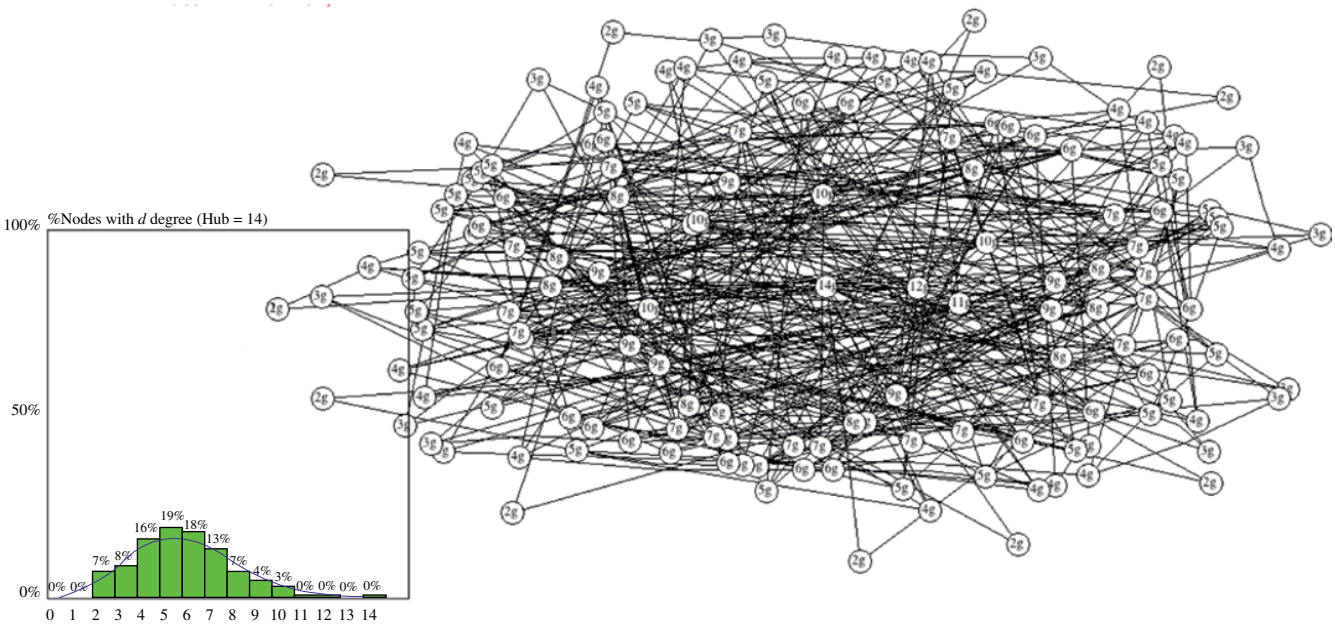


FIGURE A.2 A random network contains nodes with a binomial degree distribution.

substations, transmission lines, and homes might be represented by a network such as the one shown in Figure A.2. This particular network is a *random network*, because links were inserted between pairs of nodes at random. That is, each node was randomly selected from the 150 nodes in the network and connected by drawing a link between them. This random process was repeated 420 times—once for each of the 420 links in the network.

The process of inserting links into a network is called *percolation*. Therefore, *random percolation* produces a random network. The number of links connecting a node to other nodes is called the *degree* of the node, *g*. Degree is a measure of connectivity. Random percolation produces a network containing nodes with one, two, three, and more links. Therefore, the degree of each node varies from 0 to some maximum number. Random percolation produces networks containing nodes with a random number of connections. Therefore, the degree distribution of a random network should obey a degree distribution histogram as shown in Figure A.2.

The network of Figure A.2 produces a binomial histogram of percentage of nodes with *g* connections. We know it is a binomial histogram because the binomial distribution function shown as a solid line in Figure A.2 closely fits the bars produced by counting the number of links attached to all nodes. In other words, the degree distribution of a random network is just like Pascal's triangle—it is shaped like a symmetrical binomial distribution. The degree distribution of a network is a kind of fingerprint that tells us what kind of network it is.

A.4 CONDITIONAL PROBABILITY

In many circumstances the likelihood of an event is influenced by a prior event. For example, the likelihood of obtaining a second head (H) after obtaining the first head, in the coin-tossing example analyzed earlier, is ½, instead of ¼, because the prior event has already happened. Similarly, the likelihood of contracting a contagious disease increases if someone nearby already has the disease. Thus, probability can be conditioned on prior events rather than independent of one another. We use a vertical bar, |, to indicate that the likelihood of a future event is conditional on a prior event. Pr(A | B) is read, "the probability of event A, conditional on event B." If event B is known to have occurred, then the likelihood of A is increased, because uncertainty is reduced.

An English Presbyterian minister named Thomas Bayes (1701–1761) invented the theory of *conditional probabilities* nearly 300 years ago. His papers were published only after his death in 1762.[2] *Bayesian probability theory* was largely ignored until recently, because it is essentially a theory of belief rather than a theory of likelihood [3]. *Bayesian probability* is now the basis of artificial intelligence theories of learning, because it deals with the realities of uncertainty and belief rather than mathematical precision and likelihood.

Propositions such as "a terrorist attack is likely" are assigned a number indicating how certain we are in the proposition's accuracy. The assigned number is a *measure of belief*: zero means we believe the proposition is false, one means we believe the proposition is true, and any number in between is the degree to which we believe the proposition. For example, 0.75 indicates a high degree of belief, while 0.25 indicates a relatively low degree of belief.

For example, in the two-coin-tossing example, after it is known that the first toss produced a head (H), what is the conditional probability that the second coin is also a head? Pr(H) = ½ is the probability of obtaining a head on any toss. And we know from the previous example that the probability of obtaining two heads in two tosses is Pr(HH) = ¼. But what is the probability of obtaining a head on the second toss, if it is known that a head was obtained on the first toss? In mathematical terms, we ask what is $\Pr(H_2 \mid H_1)$, where H_1 means a head was obtained on the first toss and H_2 means a head is anticipated on the second toss. $\Pr(H_2H_1)$ is a measure of how much we believe that two heads will occur in two tosses, and $\Pr(H_2 \mid H_1)$ is a measure of how much we believe that two heads will occur in two tosses if we know for certain that the first toss produced a head. Knowledge of the first head, H_1, reduces the uncertainty of two heads in a row by $\Pr(H_1)$:

$$\Pr(H_2|H_1) = \frac{\Pr(H_2H_1)}{\Pr(H_1)} = \frac{1/4}{1/2} = 1/2$$

In general, Bayes' theorem says that the conditional probability of event A given B, Pr(A | B), is the probability of both happening, Pr(A .and. B), reduced by the certainty of Pr(B) as follows:

$$\Pr(A|B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)}$$

Bayes noted that Pr(A .and. B) is also a conditional probability, Pr(B | A)Pr(A), so the workable theorem is

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

For our purposes it will be easier to use a tabular model in place of the Bayes' formulas. Tabular modeling is illustrated in Figure A.3. For example, Figure A.3a replicates the two-coin-tossing experiment studied earlier using Pascal's combinatorial mathematics. One box represents

[2] http://en.wikipedia.org/wiki/Thomas_Bayes

(a)

	No	Yes
Toss 1 = H	0.5	0.5

	No	Yes
Toss 2 = H	0.5	0.5

Two tosses: Toss 1	Toss 2	No	Yes
T	T	1.0	0.0
T	H	1.0	0.0
H	T	1.0	0.0
H	H	0.75	0.25

(b)

	No	Yes
Toss 1 = H	0.45	0.55

	No	Yes
Toss 2 = H	0.45	0.55

Two tosses: Toss 1	Toss 2	No	Yes
T	T	1.0	0.0
T	H	0.90	0.10
H	T	0.95	0.05
H	H	0.00	1.0

FIGURE A.3 A tabular model of Pascal's ideal world of mathematical precision assumes coins are perfectly balanced so that heads and tails occur equally often. A similar tabular model of Bayes' historical world of evidence measures the number of times heads and tails occur. (a) Pascal's tabular model of tossing a coin twice and obtaining two heads (HH). (b) Bayes' tabular model of tossing an unfair coin twice and obtaining two heads (HH).

toss 1 and a second box represents toss 2. The results of these tosses flow into the two tosses box that enumerates all possible combinations of two tosses. Assuming a fair coin, the probability of obtaining a head on each toss is 0.5. Therefore, the probability of obtaining two heads in two tosses is (0.5)(0.5) = ¼ as before. But in Figure A.3a, all of the combinations and their probabilities of producing HH are listed in the two tosses box. The probability of two heads, Pr(HH), is 0.0 everywhere except for the final row where toss 1 yields H and toss 2 also yields H. This is the same result obtained earlier using Pascal's ideal world of mathematical precision.

But what happens if real-world evidence does not match with Pascal's ideal world? Suppose the number of times heads and tails are actually observed in thousands of tosses is uneven. Figure A.3b shows a Bayesian model of beliefs obtained by observation. In this scenario, H is observed 55% of the time. Toss 1 and toss 2 turn up heads 55% of the time. This chances the results. Instead of Pr(HH) = 0.25, we believe Pr(HH) = (0.55)(0.55) = 0.3025 or 32.25%.

Now suppose further observations suggest further aberration as indicated by the probabilities in the two tosses table. When both tosses turn up tails, gamblers believe the two tails appeared. But when the first toss is a tail and the second a head, gamblers believe two heads appeared 10% of the time. That is, Pr(TH) = 0.10. Similarly, gamblers believe Pr(HT) = 0.05 as shown in the two tosses table of Figure A.3b. Finally, gamblers believe 100% in Pr(HH).

So now the question is, "what is the probability of two heads occurring when an unbalanced coin is tossed two times?" To find the answer, we must expand the conditional probabilities using Bayes' chain rule. For each belief that two heads occurred given in the two tosses table, work backward through the table to obtain the conditional probabilities using the "facts" stored in all three tables:

$$\begin{aligned}\Pr(\mathrm{HH}) &= 0.0\Pr(\mathrm{T}|\mathrm{T})\Pr(\mathrm{T}) + \\ &\quad 0.10\Pr(\mathrm{T}|\mathrm{H})\Pr(\mathrm{H}) + \\ &\quad 0.05\Pr(\mathrm{H}|\mathrm{T})\Pr(\mathrm{T}) + \\ &\quad 1.0\Pr(\mathrm{H}|\mathrm{H})\Pr(\mathrm{H}) \\ &= 0.10(0.55)(0.45) + 0.05 \\ &\quad (0.45)(0.55) + (0.55)(0.55) \\ &= 0.02475 + 0.01238 + 0.3025 \\ &= 0.3395 \\ &= 33.95\%\end{aligned}$$

When the coin is fair and balanced, two heads occur with probability 25%; when the coin is more likely to turn up heads, two heads occur with probability 30.25%; and when evidence exists that on rare occasions HT and TH combinations are (erroneously) recorded as HH, two heads are recorded with probability 33.95%. As uncertainty is reduced, degree of belief in two heads increases.

A.5 BAYESIAN NETWORKS

If we can somehow combine various conditional probabilities into a system or model of beliefs, we can use the model to calculate the degree of belief that future events are eminent. This is the idea of a *Bayesian network* (BN). The BN model becomes a reasoning system rather than a rigid estimate of probability. We input initial beliefs as probabilities, run the model to see if it is predictive, and adjust the inputs as more data is acquired and uncertainty reduced.

A BN contains propositions (nodes) and their influence (links) on one another as in Figure A.4. If proposition B influences proposition A, then a link connects B to A, and we say B is the parent and A is the child. In Figure A.4a proposition Surveillance is the parent of Fertilizer and Attack. The

links define conditionality relationships between parent and child propositions: conditionality flows through a link from parent to child. (Conditionality can work both ways, depending on the application of the BN.) Conditionality for each proposition is stored as a *conditional probability table*, CPT as illustrated in Figure A.4b and c.

A BN "executes" by "reasoning" about the propositions represented as nodes. The method of reasoning is based on Bayes' theorem:[3]

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(B) + \Pr(B|\text{notA})\Pr(\text{notA})}$$

The left-hand side, Pr(A | B), is the *posteriori probability* of event A, conditional on event B. The right-hand side terms, Pr(B | A), Pr(B), and their complements Pr(B | notA) and Pr(notA), are *priori probabilities* based on historical evidence. Therefore, Pr(A | B) is conditional on evidence of Pr(B | A), Pr(A), and complements Pr(B | notA) and Pr(notA). The product Pr(B | notA)Pr(notA) is the likelihood of a *false positive*, because it is the likelihood that B is true when its precursor A is not.

A.6 BAYESIAN REASONING

This kind of machine reasoning combines evidence-based logic with probability theory. As conditional probabilities of prior events become known from accumulated evidence, uncertainty in BN propositions declines, yielding more "belief" in the proposition. Bayes' theorem treats probabilities as evidence—a confusing departure from Pascal's interpretation of probability—so a more thorough example is given here to clarify the significance of Bayes' work.

Consider the elementary BN of Figure A.4. Suppose a law enforcement agency (LEA) wants to estimate the probability of a terrorist attack given evidence obtained through *suspicious activity reports* (SARs). The agency begins by building a model of a typical terrorist bombing incident as shown in Figure A.4. Historical SARs provide evidence that bombers often visit their target site several times before an attack. In addition, historical evidence suggests that terrorists have made bombs from fertilizers, so the LEA also tracks fertilizer purchases. How are these "facts" used to calculate threat?

The relationship between surveillance, fertilizer purchases, and attacks is represented in Figure A.4a as a network consisting of Surveillance, Fertilizer, and Attack nodes. These nodes represent propositions—unsubstantiated claims—with associated degrees of believability. Their "truth" is questionable. Initial estimates of likelihood are mere guesses, but these guesses should get better as more evidence is used to update the "truthness" of one or more of the propositions.

[3]http://en.wikipedia.org/wiki/Bayes'_theorem

Recall that conditionality is represented by links from parent-to-child nodes. In the illustrative example of Figure A.4, Surveillance is the parent of both Fertilizer and Attack nodes, and Fertilizer is the parent of Attack. Truth, or the degree of belief, is transmitted through the network via these links. Therefore, a change in one proposition node propagates to others. Truth emerges as a by-product of this propagation.

Figure A.4a–c shows the priori estimates of the likelihood that a suspect will visit a target many times (5%) and the likelihood that the suspect will buy fertilizer after visiting the target (90%). Every proposition has an output value (true or false, Yes or No, Buy or None) that is conditional on its input values. Therefore, if a proposition has one input link that can be either true or false, it must have a Bayesian estimate of the output for each of the possible inputs. If a proposition has two inputs, each with a possibility of being true or false, then the output depends on four cases: TT, TF, FT, and FF, where T = True and F = false. (The combinations can also be Yes, No or On, Off or Buy, Don't Buy as well.) Figure A.4c shows all possible combinations of input values for the Attack proposition, which is conditional on two links with values of None, None; None, Yes; Buy, None; and Buy, Yes. Note that these probabilities must sum to 1.0 across every row of the CPT in each proposition.

Initially, the BN represents reality as users perceive it. For example, the initial likelihood of a terrorist visiting a target several times before attacking is assumed to be 5% (see Fig. A.4a). This is merely a belief and is unsubstantiated without evidence. It is interesting to note that these estimates need not be especially accurate, because their impact on the final answer will be altered as new evidence comes in and is incorporated into the BN. This "fuzziness" is one of the major advantages of using Bayesian belief networks in place of subject matter experts, alone.

Figure A.4d and e shows how new evidence is used to update the BN and therefore increase the believability of an eminent attack. The network is updated as situational awareness reports arrive and are incorporated. A new estimate of the posteriori probability of an attack is automatically recalculated. (Typically, a BN software application is used to perform these calculations.)

Suppose an SAR indicates an unusual interest in a certain target by a suspect. The priori probability of Surveillance can now be changed from 5 to 100%, because of the new evidence. The certainty created by the new evidence increases the posteriori threat to 91.6% (see Fig. A.4d). Reducing uncertainty in one part of the BN increases our belief in an attack in another part of the BN. When a subsequent SAR indicates that the same suspect has purchased a large quantity of fertilizer, the Fertilizer node is updated to 100%, and the BN automatically recalculates the posteriori attack probability (see Fig. A.4e). As more evidence is gathered, uncertainty is further reduced, and the

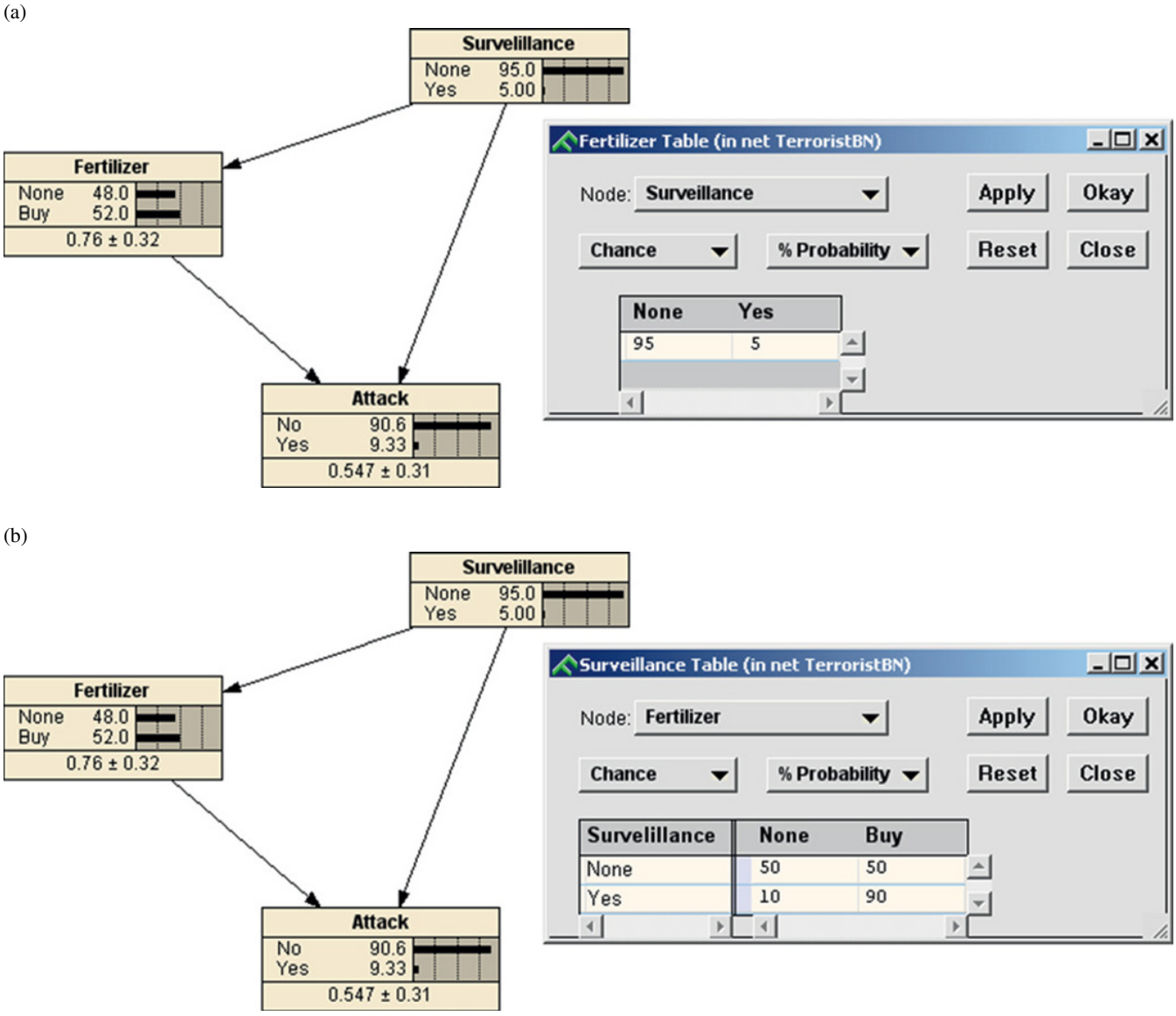


FIGURE A.4 A Bayesian network (BN) model of threat. (a–c) Prior beliefs of a suspected terrorist performing surveillance on a potential target; likelihood of buying fertilizer to build a bomb, conditional on precursor surveillance; and likelihood of an attack, conditional on surveillance and buying fertilizer. (d–e) The increase in posteriori probability of an attack after it is certain that the suspect has performed surveillance and then purchased large amounts of fertilizer. *Netica*, from Norsys Software, was used to build the Bayesian belief network and perform the calculations illustrated here. (Norsys Software Corp., Vancouver, Canada, V6S 1K5.) (a) Probability of surveillance as a precursor to an attack. (b) Probability of buying fertilizer conditioned on surveillance: 0.48 = 0.50(0.95) for None, and 0.52 = 0.50(0.95)+0.90(0.05) for Buy.

posteriori probability increases from 9.33 to 91.6% and finally 99%.

BN theory is a tool for calculating *posteriori* probabilities—it attempts to predict the future from the past. Unlike static and rigidly determined probabilities, Bayesian probabilities are beliefs containing uncertainty. But a BN is only a model of what we believe to be true about the real world when the data contains uncertainty. A proposition is more likely to be true if its degree of belief is high, but keep in mind that these estimates are only as good as the model and input data.

Bayesian belief networks resonate with Laplace's rule of succession, because both theories incorporate doubt in their model of reality. Laplace squeezes out lingering uncertainty using overwhelming historical evidence. Bayes squeezes out uncertainty using convincing evidence and conditional probability. BN are, however, based on sound (mathematical) principles—Bayes' theorem provides a mechanical method of expressing the amount of uncertainty reduction that is made possible by incorporating more information.

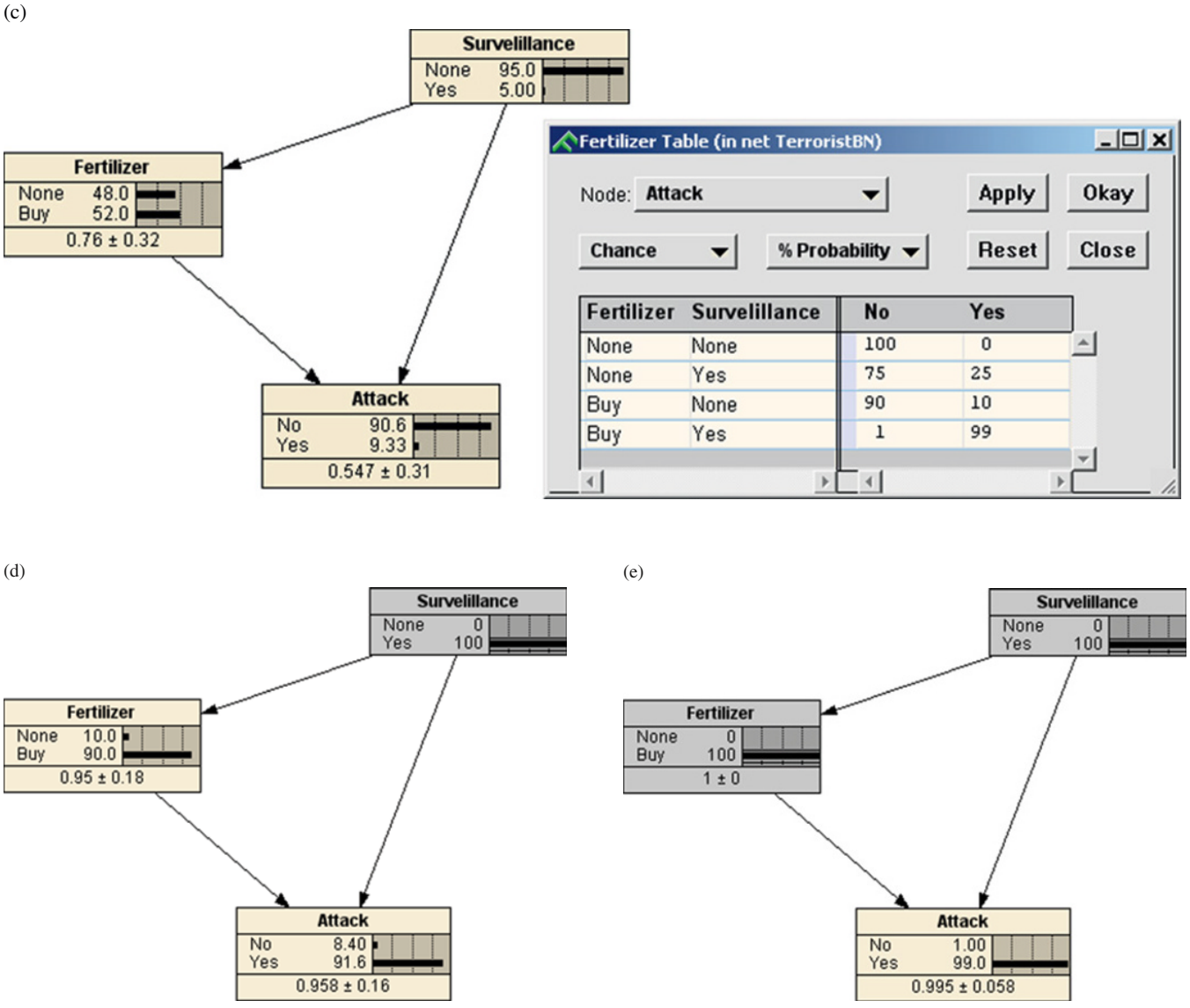


FIGURE A.4 (Continued) (c) Probability of a terrorist attack conditioned on surveillance and buying fertilizer is 9.33% = 0.25 Pr(Yes | None .and. Yes) + 0.10 Pr(Yes | Buy .and. None) + 0.99 Pr(Yes | Buy .and. Yes) = 0.25(0.10) Pr(Surveillance = Yes) + 0.10(0.50) Pr(Surveillance = None) + 0.99(0.9 0) Pr(Surveillance = Yes) = 0.25(0.10)(0.05) + 0.10(0.50)(0.95) + 0.99(0.90)(0.05) = 0.0933. (d) Probability of attack after it is known that suspect has performed surveillance on the target. Note the degree of belief that an attack will occur rises from 9.33 to 91.6%. (e) Posteriori probability of attack after it is known that suspect has performed surveillance and bought fertilizer. Note the degree of belief in an attack rises from 91.6 to 99.0%.

Unfortunately, the knowledge required to build and operate a BN may exceed the capabilities of an agency or risk assessment operator. BN construction requires a combination of subject matter expertise and facility with Bayes' theorem and corresponding modeling tools. Fortunately, a number of software packages exist to do the calculations once a BN is constructed.[4] But someone must customize each model for each situation. I used Norsys Software's Netica to illustrate BN modeling in Figure A.4, which made it possible to build a model without knowing the math.

[4]http://www.cs.ubc.ca/~murphyk/Bayes/bnsoft.html

REFERENCES

[1] Boyer, C. B. *A History of Mathematics*, 2nd ed, New York: John Wiley & Sons, 1991.

[2] Bailey, R. Don't Be Terrorized. Reason.com, August 11, 2006. Available at http://reason.com/archives/2006/08/11/dont-beterrorized. Accessed June 25, 2014.

[3] Bellhouse, D. R. The Reverend Thomas Bayes: A Biography to Celebrate the Tercentenary of his Birth, *Statistical Science*, 19, 1, 2004, pp. 3–43.

FURTHER READING

Bernstein, P. L., *Against the Gods: The Remarkable Story of Risk*, New York: John Wiley & Sons, 1996. Starting point for anyone wanting to understand risk.

Boyer, C. B., *A History of Mathematics*, 2nd ed, New York: John Wiley & Sons, 1991. The authoritative book on mathematics traces human thought and its formalization as modern mathematics in the twentieth century, going back to the Ancient Greeks.

Grossi, P. and Kunreuther, H. *Catastrophe Modeling: A New Approach to Managing Risk*, New York: Springer, 2005, pp. 245. Support for the exceedence probability approach to risk assessment with examples of its practical application.

Hanson, R., Catastrophe, Social Collapse, and Human Extinction, in *Global Catastrophic Risks*, ed. M. Rees, N. Bostrom, and M. Cirkovic, Oxford/New York: Oxford University Press, July 17, 2008, pp. 363–377. An interesting non-technical exploration of catastrophic risk with lots of supporting data.

Perrow, C., *Normal Accidents*, Princeton: Princeton University Press, 1999, pp. 450. Pioneering book that established the earliest theory on catastrophes, disasters, and failure in human-made systems. This is an essential read for anyone contemplating catastrophe theory.

Ramo, J. C., *The Age of the Unthinkable*, New York: Little, Brown & Company, 2009, pp. 280. Popular tome on complex adaptive systems theory as it might apply to socio-economic-political systems. Ramo is another fan of Per Bak's work.

Taleb, N. N., *Fooled by Randomness*, New York: Random House, 2005, pp. 316. Entertaining and erudite treatise on how humans mistake randomness for cause-and-effect.

U.S. Department of Homeland Security, *National Infrastructure Protection Plan*, U.S. Department of Homeland Security, Washington, DC, 2009, pp. 100. Document setting forth the Department of Homeland Security's strategy of riskoriented decision-making, and includes DHS's intuitive approach to risk and resilience.

APPENDIX B

MATH: RISK AND RESILIENCE

Probability theory was devised largely for a very practical reason—gambling. Predicting the amount of money one could make by "risking" capital at the card table and roulette wheel was probability theory's "killer app." A handsome profit can be had by correctly predicting the future outcome of a game of chance. Indeed, study of this killer app accelerated for the next 200 years and continues today. Edward Oakley Thorp (1932–)—an American mathematics professor, author, hedge fund manager, and blackjack player best known as the "father of the wearable computer"—demonstrated perhaps the most dramatic application of probability theory to gambling in 1961. Thorpe used a concealed computer to beat the blackjack tables in Las Vegas. He documented his technique in a best seller titled *Beat the Dealer* in 1962. (His technique is the famous card counting method.)

Thorpe was following in the footsteps of Geronimo Cardano (1501–1576)—a famous Milanese physician. More importantly, he was also a compulsive gambler, earning the name "Gambling Physician." Gambling drove Cardano to formulate early ideas that later became the basis of modern risk assessment. He combined probability estimates with gains and losses—consequences—to formulate the early idea of risk. He was concerned with predicting how much money might be made by repeatedly playing a certain game and, on the downside, how much money might be lost. Cardano intuitively understood risk as his *expected gain or loss* after a hard day of gambling.

Daniel Bernoulli (1700–1782)—a third-generation grandson of the famous family of Swiss mathematicians—formalized Cardano's intuition 200 years later. Bernoulli's risk equation is the foundation of modern *expected utility theory* (EUT) (1738). According to Bernoulli, risk is the product of the probability of a certain outcome and its consequence: $R = Pr(C)C$, where $Pr(C)$ is the probability of losing C dollars, say, and C is the loss measured in dollars. When n independent events are possible, risk is simply the sum of all expected values: $Pr(C_1)C_1 + Pr(C_2)C_2 + \cdots + Pr(C_n)C_n$. This breakthrough in risk calculation continues to be used today in financial and engineering calculations.

Bernoulli's formulation established the field of a priori probability based on the simple observation that the likelihood of an event is the ratio of number of ways the event can happen to the total number of events possible. The total number possible is the space of all events, while the number of ways a certain event can occur is a subset of the space. For example, the space of events for a tossed coin is [H, T], representing heads or tails. The size of this space is 2. If we want to know the a priori probability of H, we form the ratio of number of ways an H can occur versus the total number of events, which is 2. Thus, the a priori probability of H is ½, and similarly the a priori probability of T is also ½, assuming the coin is balanced and fair.

The Bernoulli formulation assumes a finite space of possible events and the enumeration of events of interest. This is in sharp contrast to a posteriori probability, which is based on observations of the past. The outcome of a tossed coin is no longer based on combinations of possible outcomes, but instead it is based on the past. Suppose a certain coin is tossed 1000 times and the number of time H occurs is recorded. Further suppose the number is 512. The a posteriori

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

probability of the next toss turning up heads is 512/1000 or 0.512 instead of 0.500.

The a priori estimate of likelihood is based strictly on mathematics, while the a posteriori estimate is based on belief established by observation. Thus, a posteriori probability is often called a belief system as opposed to a probabilistic system. In fact, belief systems are as old as a priori probability going all the way back to Presbyterian reverend Thomas Bayes (1701–1761), an English statistician who invented a posteriori probability that now bears his name. Bayesian probability is based on belief established by historical record and observation. It assumes that the future is based on the past and that belief in the outcome of an event such as coin tossing is directly proportional to the number of times an event has occurred in the past.

Bayesian probability is particularly useful in predicting the likelihood of events that are related. For example, a cloudy day is more likely to rain than a clear day. As clouds mount, so does the probability of rain. When two or more events are related in this way, they form a network of related events. Rain events are linked to cloud events. As the probability of clouding increases, the probability of rain increases. Thus, the probability of rain is conditional on the probability of cloudiness. Bayesian networks (BNs) are models of conditional probability.

The formulation of risk in this book is a simple example of conditional probability in the sense of Bayesian probability. Risk is the product of TVC, where V is the conditional probability of an asset failing given it is threatened. In this sense, the fault trees described here are simple BNs containing a priori and conditional probabilities. The risk equation R = TVC is a hybrid of Bernoulli's and Bayes' thinking about probability.

B.1 EXPECTED UTILITY THEORY

EUT is the modern basis of *risk-informed decision-making* used to decide how best to allocate resources to either increase expected gains or reduce expected losses. Reducing expected losses is the objective of CIP, so we will focus on it. But there are several alternative approaches to risk reduction that will be surveyed here. A rough breakdown of the surveyed approaches only scratches the surface:

PRA and fault trees
Bayesian belief networks
Game theory
Network risk and resilience

B.1.1 Fault Trees

Figure 2.1 is used as an illustrative example here. In addition, threat T, vulnerability V, and consequence C are assumed to define risk as follows:

$$\text{Risk} = \text{T}(\text{attacked})\,\text{V}(\text{successful if attacked})\,\text{C}(\text{failure})$$

Or in simple algebraic terms, R = TVC. The following mathematical models underpin the model-based risk analysis (MBRA) fault tree software, which is used throughout this book.

Fault trees represent threat–asset pairs obtained by considering hazards: $h_1, h_2, h_3, \ldots, h_k$ paired with an asset that is attacked with threat probability $t_1, t_2, t_3, \ldots, t_k$. The threat–asset pairs are placed into a fault tree connected by logic relations: AND, OR, or XOR. Thus a fault tree is a logic model of the likelihood of asset failure when one or more of its components fail. In this case, the risk associated with each threat–asset pair, $r_i = t_i v_i c_i$, and the likelihood of the entire fault tree failing is a Boolean expression obtained by tracing the flow of one or more faults from one or more hazards to the root of the fault tree. Probability and logic are combined using De Morgan's laws and simple relationships in probability theory. In this way, the risks of each threat–asset pairs are combined into a system of risks that represent the various failure modes of the asset. This idea will be illustrated with the broken car fault tree of Figure 2.1.

MBRA defines risk in fault trees as the total expected loss from all threat–asset pairs:

$$R = \sum_i^n t_i v(p_i) c_i$$

$$v(p_i) = v_i(0) e^{-\gamma_i p_i}; \quad \gamma_i = -\ln\left[\frac{v_i(\infty)}{v_i(0)}\right] / p_i(\infty)$$

For each threat–asset pair, *i*:

- t_i : probability of hazard or attack
- $v(p_i)$: the probability of destruction vice investment p_i
- p_i : investment to reduce vulnerability from initial value of $v_i(0)$
- $p_i(\infty)$: investment that reduces vulnerability to elimination cost $v_i(\infty)$
- $v_i(\infty)$: vulnerability after an investment of $p_i(\infty)$
- $v_i(0)$: initial vulnerability of asset or component
- c_i : consequence of a successful attack on asset or component
- n : number of threat–asset pairs in the fault tree

The risk data is constrained by a budget, P:

$$P = \sum_{i=1}^n p_i$$

$p_i \geq 0$: P is total budget available to reduce vulnerability, given

Fault tree vulnerability, or probability that the entire fault tree fails, is obtained by working up the fault tree from threat blocks at the lowest level to the root of the fault tree while applying the following equations for propagated vulnerability.

Hazards connected to a component by an AND gate produce a failure in the component with probability defined by the product of vulnerability values, OR gates with probability defined by De Morgan's law, and XOR gates with probability defined by the sum of exclusive failure event probabilities. The equations for components connected to other components or threats by one of the logic gates are:

AND gate

$$\Pr(\text{AND}) = \prod_{i \in \text{component}} t_i v_i (p_i)$$

OR gate

$$\Pr(\text{OR}) = 1 - \prod_{i \in \text{component}} \left[1 - t_i v_i (p_i)\right]$$

XOR gate

$$\Pr(\text{XOR}) = \sum_{s \in \text{component}} \left\{ t_s v_s (p_s) \prod_{j \neq s} \left[1 - t_j v_j (p_j)\right] \right\}$$

For example, using the data in Table 2.1a, the following initial risks and fault tree vulnerability are obtained:

$$\begin{aligned}
&t_1 = 0.50; v_1 = 0.50; t_1 v_1 = 0.25 \\
&t_2 = 0.80; v_2 = 0.50; t_2 v_2 = 0.40 \\
&t_3 = 0.25; v_3 = 1.00; t_3 v_3 = 0.25 \\
\Pr(\text{AND}) &= t_1 v_1 \cdot t_2 v_2 \cdot t_3 v_3 = (0.25)\cdot(0.40)\cdot(0.25) \\
&= 1/40 = 0.0250 \\
\Pr(\text{OR}) &= 1 - (1 - t_1 v_1)(1 - t_2 v_2)(1 - t_3 v_3) \\
&= 1 - (0.75)(0.60)(0.75) = 0.0375 \\
\Pr(\text{XOR}) &= 0.25(1 - t_2 v_2)(1 - t_3 v_3) + 0.4(1 - t_1 v_1)(1 - t_3 v_3) \\
&\quad + 0.25 \cdot 1(1 - t_1 v_1)(1 - t_2 v_2) = 0.450
\end{aligned}$$

Note that XOR produces the highest vulnerability, OR the next highest, and AND produces the lowest vulnerability. Why? AND multiplies TV values together, which produces the lowest possible values. OR fault trees adds together all 2^n possible combinations, which makes the OR larger than the AND tree vulnerability. XOR is the most interesting because it considers the least number of combinations—only n. But because an XOR tree excludes all but one threat–asset pair at a time, it produces the highest vulnerability.

The initial risk associated with the fault tree in this example is simply the sum of initial risks across all hazards:

$$\text{R} = \sum_{i=1}^{3} t_i v_i c_i = (0.25)(300) + (0.40)(300) + (0.25)(300) = 270$$

But this expression is inaccurate for the XOR fault tree, because the probability of each threat–asset pair occurring must also include the probabilities that the other two pairs do not occur. Therefore, the expression for an XOR fault tree is

$$\begin{aligned}
\text{R}(\text{XOR}) &= \sum_{s \in component} \left\{ t_s v_s (p_s) \prod_{j \neq s} \left[1 - t_j v_j (p_j)\right] c_s \right\} \\
&= 0.25(1 - t_2 v_2)(1 - t_3 v_3)\cdot 300 + 0.4(1 - t_1 v_1)(1 - t_3 v_3)\cdot 300 \\
&\quad + 0.25 \cdot 1(1 - t_1 v_1)(1 - t_2 v_2)\cdot 300 \\
&= 0.25(0.6)(0.75)\cdot 300 + 0.4(0.75)(0.75)\cdot 300 \\
&\quad + 0.25(0.6)(0.75)\cdot 300 \\
&= (0.1125 + 0.2250 + 0.1125)\cdot 300 = 0.45 \cdot 300 = 135
\end{aligned}$$

The risk of an XOR fault tree is one-half as much as the risk of the other two types of fault trees. Why? The probability of a hazard not occurring reduces the TV values of every threat–asset pair in the tree. This aspect of an XOR tree also makes it much more difficult to calculate the optimal allocation of resources to minimize risk.

B.1.2 Fault Tree Minimization

Resource allocation asks, "What is the best allocation of a fixed budget P threat–asset pairs such that risk is minimized? Optimal allocation of P to n threat–asset pairs in a fault tree with only AND and OR gates is directly calculated from the risk equation:

$$\min_P \langle \text{R} \rangle = \sum_{i=1}^{n} t_i v_i (p_i) c_i v$$

$$v_i (p_i) = v_i (0) \exp(-\gamma_i p_i); \quad \gamma_i = \frac{-\ln\left[\dfrac{v_i(\infty)}{v_i(0)}\right]}{p_i(\infty)}$$

$$\text{P} = \sum_{i=1}^{n} p_i; \quad p_i \geq 0$$

$$v_i (\infty) \approx 0.05$$

Typically, a small value such as 0.05 or 0.01 is used in place of zero, because ln[0] is meaningless, and MBRA assumes it is impossible to remove all vulnerability.

Minimization is done by classical optimization using a Lagrange multiplier λ:

$$\ln(\lambda) = \frac{\sum_{i=1}^{n} \dfrac{\ln(\gamma_i t_i v_i(0))}{\gamma_i} - \text{P}}{\sum_{i=1}^{n} \dfrac{1}{\gamma_i}}$$

$$p_i = \frac{\ln(\gamma_i t_i v_i(0))}{\gamma_i} - \ln(\lambda)$$

These equations work for a fault tree with AND and OR gates and provides a starting point for optimizing a fault tree with one or more XOR gates. But they fail to calculate the true minimum because there is no known closed-form formula for minimizing an XOR fault tree. Therefore, an iterative numerical method is used to approximate the minimum.

B.1.3 XOR Fault Tree Allocation Algorithm

1.0. Temporarily replace XOR gates with OR gates and find an initial allocation of P to $[p_1, p_2, \ldots, p_n]$. Revert to the original fault tree with its XOR gates.

2.0. Repeat for $t = 1, 2, \ldots$

2.1. Select a donor threat-asset at random, DONOR.

2.2. Select a target threat-asset at random, TARGET

2.3. Save p_{DONOR} and p_{TARGET}, as SAVED_DONOR, SAVED_TARGET.

2.4. Calculate and save fault tree risk, RISK.

2.4. Add a random amount, STEP, to p_{TARGET}

2.5. Subtract STEP from p_{DONOR}

2.6. Recalculate risk, NEW_RISK

2.7. if NEW_RISK > RISK

2.7.1. Restore p_{DONOR} and p_{TARGET} to saved SAVED_DONOR and SAVED_TARGET

2.8. Until there is no additional change to RISK, e.g. $\left|\frac{\text{RISK}(t) - \text{RISK}(t-1)}{\text{RISK}(t)}\right| < \varepsilon$

Fault tree optimization on trees of 20–30 threat–asset pairs takes a matter of seconds on a desktop computer using this algorithm. Note that convergence to a true minimum risk is not guaranteed, but as a practical matter, allocations are very close. Additionally note that the STEP size is a random number, typically calculated as a random fraction of the difference between allocation p_i and $p_i(\infty)$.

B.2 BAYESIAN ESTIMATION

The second major criticism of PRA concerns the placement of T on the right-hand side of the risk equation. Critics say that T should be an *output* rather than an input to risk assessment. That is, threat should be a by-product of risk assessment, because terrorists are more likely to attack weaker targets than stronger or better protected targets. According to the critics of PRA, a rational terrorist might attack the most vulnerable target to maximize his or her expected utility. Alternatively, a terrorist might simply take advantage of an opportunity and ignore rationality. In either case, the notion of a fixed value for T is considered faulty logic.

Two alternatives to EUT are presented. The Bayesian Network (BN) approach uses evidence of both capability and intent to predict a terrorist activity. In the second alternative, game theory is used to optimize both terrorist and defender allocations. In the game-theoretic approach, both terrorist and defender attempt to optimize their respective objectives—the terrorist wants to maximize risk, while the defender wants to minimize risk. But first, what is a Bayesian belief network?

B.2.1 Bayesian Networks

Bayes defined probability as a *belief* rather than a frequency. Propositions such as "a terrorist attack is likely" are assigned a number indicating how certain we are in the proposition's veracity. The number is a *measure of belief*: zero means we believe the proposition is false; one means we believe the proposition is true, and any number in between is the degree to which we believe the proposition. For example, 0.75 indicates a high degree of belief, while 0.25 indicates a relatively low degree of belief.

If we can somehow combine various propositions into a system or model of beliefs, we can test the model against actual data. This is the idea of a BN. The BN model becomes a reasoning system. We input initial beliefs as probabilities, run the model to see if it is predictive, and adjust the inputs as more data is acquired and uncertainty is reduced.

A BN contains propositions (nodes) and their influence (links) on one another as in Figure 2.6. If proposition S influences proposition B, then a link connects S to B, and we say S is the parent and B is the child. In Figure 2.6 proposition S (Surveillance) is the parent of B (Bomb) and A (Attack Bridge). The links define conditional relationships between parent and child propositions: conditionality flows through a link from parent to child. (Conditionality can work both ways, depending on the application of the BN.) Conditionality is stored in each proposition as a *conditional probability table* (CPT) as shown in Figures 2.6.

A BN "executes" by "reasoning" about the propositions represented as nodes. The method of reasoning is based on Bayes' theorem[1]:

$$\Pr(\text{A} \mid \text{B}) = \frac{\Pr(\text{B} \mid \text{A})\Pr(\text{A})}{\Pr(\text{B} \mid \text{A})\Pr(\text{B}) + \Pr(\text{B} \mid \text{notA})\Pr(\text{notA})}$$

The left-hand side, Pr(A | B), is the *posteriori probability* of event A, conditional on event B. The right-hand side terms, Pr(B | A), Pr(B), and their complements Pr(B | notA) and Pr(notA) are *priori probabilities* based on beliefs—more accurately, *degree of belief*s—a measure of how confident we are that a certain fact is actually true. Therefore, Pr(A | B) is

[1] http://en.wikipedia.org/wiki/Bayes'_theorem

conditional on (historical) evidence of Pr(B | A), Pr(A), and complements Pr(B | notA) and Pr(notA). The product Pr(B | notA)Pr(notA) is the likelihood of a *false positive*, because it is the likelihood that B is true when its precursor A is not.

For our purposes, a different form of Bayes' theorem is used—the chain rule for calculating the degree of belief at the end of a chain of conditionals:

$$\Pr(A \mid B,C) = \sum_{B}\sum_{C} \Pr(A \mid B)\Pr(B \mid C)$$

This notation quickly becomes too cumbersome to make sense of the BN, so a CPT is used to store all possible combinations and their probabilities. The CPTs of Figure 2.6 are obtained from evidence—historical data or from experience. Once entered into a CPT, they are used to answer questions such as "what is the probability of an attack?" Even after they are entered, additional evidence may indicate a change—perhaps and improvement in accuracy or new information—so the CPT values may be changed.

Consider the BN of Figure 2.6. How are these propositions and CPTs used to calculate threat? Threat is equivalent to Pr(A = Yes | B, S) over all possible values of B and S. Figure B.1 expresses the calculations as a tree of all conditional probabilities involved in computing Pr(A = Yes | B, S) for all combinations of B and S. It also shows how the threat value 0.667 was obtained for the BN of Chapter 2. The conditional probabilities along each vertical branch of the tree are multiplied together and then summed across all four branches to obtain 0.667.

The relationship between surveillance, bomb-building capability, and attacks is represented in Figure B.1 as a tree consisting of conditional probabilities taken from the CPTs of Chapter 2. Change one entry in the CPTs and the entire tree of calculations must be repeated to arrive at a new T.

BN theory is a tool for calculating *posteriori* probabilities–it attempts to predict the future from past evidence. Unlike static and rigidly determined probabilities, Bayesian probabilities are beliefs containing uncertainty. But a BN is only a model of what we believe to be true about the real world when the data contains uncertainty. A proposition is more likely to be true if its degree of belief is high, but keep in mind that these estimates are only as good as the model and input data.

Unfortunately, the knowledge required to build and operate a BN may exceed the capabilities of an agency or risk assessment operator. BN construction requires a combination of subject matter expertise and facility with Bayes' theorem and corresponding modeling tools. Fortunately, a number of software packages exist to do the

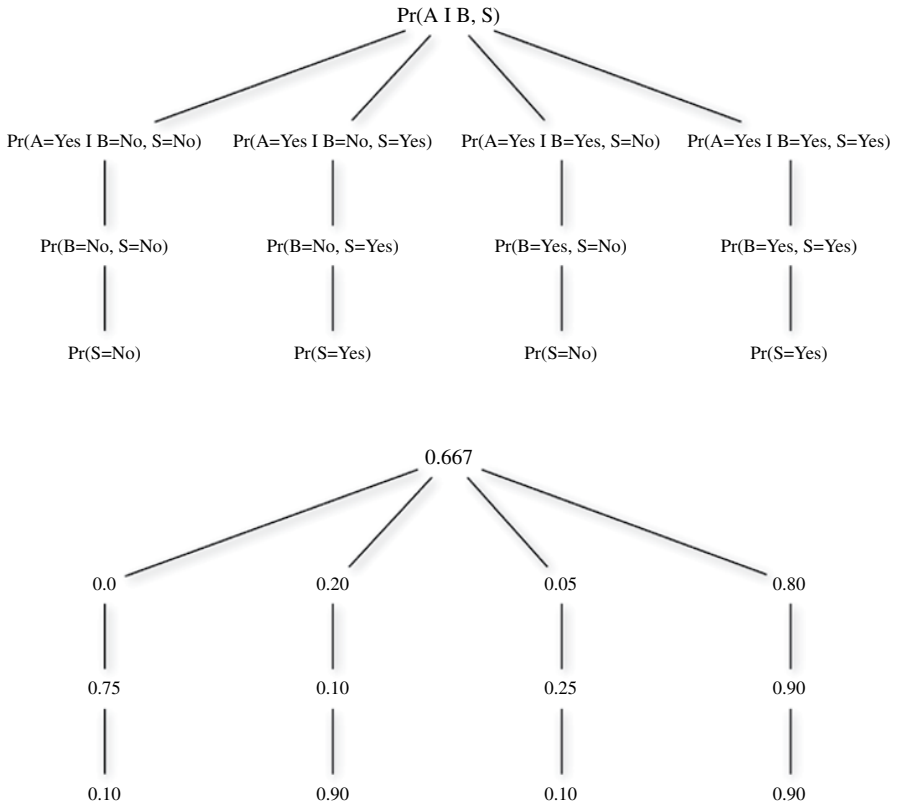


FIGURE B.1 The computation tree of the BN in Chapter 2 contains all possible combinations of Yes and No answers to each proposition in the BN.

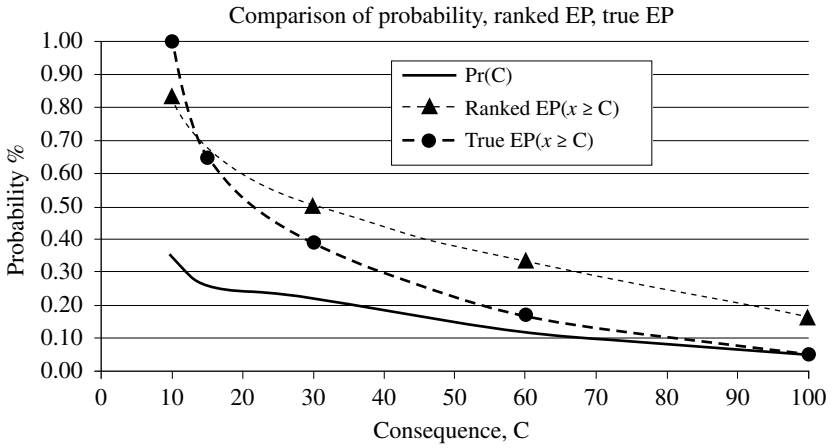


FIGURE B.2 Comparison of probability distribution, ranked exceedence probability, and true exceedence probability illustrates the differences.

calculations once a BN is constructed.[2] But someone must customize each model for each situation. I used *Norsys Software's Netica* to illustrate BN modeling, which made it possible to build a model without knowing the math.

B.3 EXCEEDENCE AND PML RISK

Exceedence probability, EP($x \geq$C), is the probability that x equals or exceeds C. It is often used by the insurance industry to estimate risk for the purpose of establishing insurance premiums, because insurance companies want to know their maximum exposure. Thus, *probable maximum risk* (PML) is defined as the product of exceedence probability and consequence:

$$\mathrm{PML}(\mathrm{C}) = \mathrm{EP}(x \geq \mathrm{C}) \cdot \mathrm{C}$$

Simple exceedence probability is obtained by ranking a set of observations x_i; $i = 1, \ldots, n$, from highest to lowest (1 is the highest, and n is the lowest), and plotting the rank of x_i versus x_i. This is called *ranked exceedence*, because it denotes the likelihood of the number of disastrous events that cause damage greater than or equal to x_i:

$$\mathrm{EP}(x_i) = \frac{\mathrm{Rank}(x_i)}{n+1}$$

Unfortunately, this definition produces the same likelihood values regardless of the underlying probability distribution, Pr(x_i). If disastrous events are of size 10, 20, and 30, say, the ranked exceedence is identical regardless of the likelihood of each event, Pr(10), Pr(20), and Pr(30). For example, if the values of Pr(x_i) for a hurricane are 50, 30, and 20% and the values for earthquakes are 25, 60, and 15%, the ranked EP values will be the same. Ranked EP does not measure probability of an event. It measures the probability of n events greater than or equal to consequence, C.

Alternatively, *true EP* is the probability of a single event greater than or equal to consequence C. It is obtained by summing Pr(x_i) from the right-hand side to the left: $i = n$, $n - 1$, …, 1, as follows:

$$\mathrm{EP}(x_i) = \sum_{j=n}^{i} \Pr(x_j)$$

The difference between ranked and true EP is illustrated in Figure B.2. First, note that true EP always starts at 100%. This means the probability of an event of size zero or greater is 100%—or conversely, the probability of an event smaller than zero is zero. Alternatively, ranked EP always starts at $1/(n+1)$ on the right and ends with $n/(n+1)$ on the far right tail. All values in between are the same, regardless of the underlying frequency distribution, Pr(C). This makes sense, because ranked exceedence counts events. In Figure B.2 the values of ranked and true EP at the extreme right end differ, because ranked EP is always $n/(n+1)$. But true EP is identical to the probability of the most extreme event.

B.3.1 Modeling EP

Most hazards produce long-tailed true EP curves as shown in Figure B.5. This says that small incidents are much more likely than large incidents. The probability of an extremely consequential event—a *black swan*—is vanishingly small. The long-tailed exceedence curve can be approximated by a power law, in most cases, which simplifies calculations. A power law of *fractal dimension q* is simply

[2] http://www.cs.ubc.ca/~murphyk/Bayes/bnsoft.html

$$\mathrm{EP}(x) = x^{-q}; \quad q > 0$$

Exponent q is called the fractal dimension because power laws are very simple self-similar fractals. As it turns out, q has an even more significant meaning when a power law is applied to the definition of PML risk:

$$\mathrm{PML} = \mathrm{EP}(x)x = \frac{x}{x^q} = x^{1-q}$$

Note that PML risk either increases or decreases as x increases, depending on the value of fractal dimension, q. This is illustrated graphically in Figure B.5 and mathematically below:

$$\lim_{x \to \infty} \left(x^{1-q}\right) = \begin{cases} \infty & q < 1 \\ 0 & q > 1 \end{cases}$$

When $q < 1$, the hazard that produced the PML curve shown in Figure B.5 is called a high-risk hazard, and when $q > 1$, the hazard is called low-risk hazard. This reason for this classification is clear in Figure B.5, because PML risk either goes up or down with increasing consequence.

B.3.2 Estimating EP from Data

How is true exceedence probability calculated? True EP is the sum of the underlying frequency distribution obtained by simulation or from historical records:

$$\mathrm{EP}(x_i) = \sum_{j=n}^{i} \Pr(x_j)$$

So the first step in calculating EP is to tabulate $\Pr(x_j)$. To illustrate this, consider the California forest fire spreadsheet shown in Table B.1 and its corresponding graph in Figure B.3. CalFire tabulates forest fires by number of acres destroyed. Therefore, consequence is reported in acres and frequency in counts. The counts are converted into a histogram by dividing the number of fires of mean consequence by the total reported. As you can see in Figure B.3, the frequency of fires of a certain size is lopsided, but not a pure power law.

TABLE B.1 This spreadsheet of 2007 raw data and EP calculations was used to produce the graphs in Figure B.3

Mean acres	Number	Frequency	EP (actual)	Power law ($q = 0.60$)	PML risk	Risk profile
1	1184	0.33	1.00	1.00	1.00	0.12
5	2091	0.58	0.67	0.38	3.36	0.40
50	262	0.07	0.09	0.10	4.64	0.56
200	39	0.01	0.02	0.04	4.04	0.49
600	25	0.01	0.01	0.02	5.65	0.68
3000	3	0.00	0.00	0.01	7.48	0.90
5000	6	0.00	0.00	0.01	8.31	1.00
Total number:	3610			Peak PML risk:	8.31	

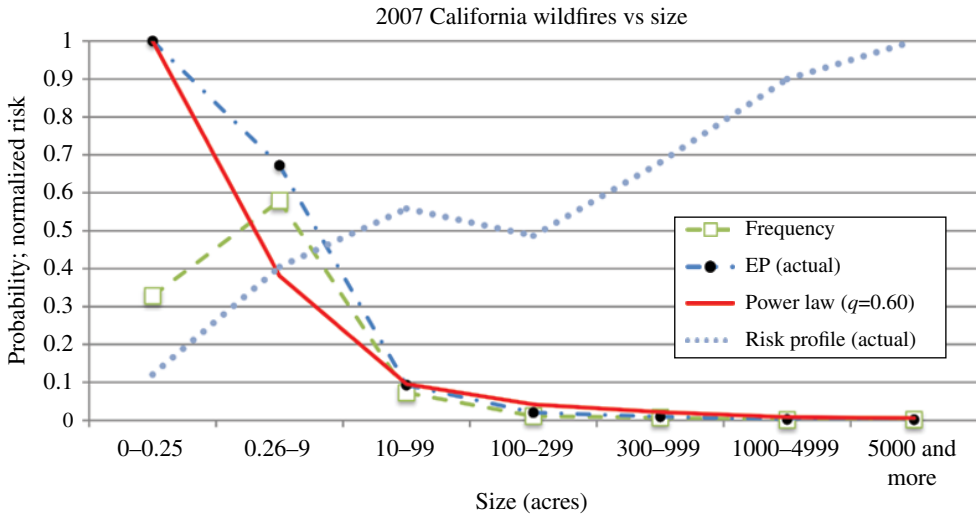


FIGURE B.3 Forest fires in Southern California are high risk according to the long-tailed exceedence probability and risk profile as shown here. The fractal dimension, $q = 0.60$, is less than 1. The exceedence probability is calculated from the frequency data.

The next step is to sum the frequency data from right to left, checking to make sure it totals 100% at the origin. The actual EP is shown in Figure B.3 as a blue dashed line with black dots. The fractal dimension is obtained by plotting log(EP) versus log(size) and fitting a straight line to the log–log plot. This calculation is not shown here, but the result is $q = 0.60$. Therefore, this is a high-risk hazard as indicated by the risk profile.

Risk is simply the product of EP and size. In other words, the vertical axis values for EP are multiplied by the horizontal axis values for size. This curve is shown as a dotted line that generally trends upward as size increases. The bigger the 2007 forest fires, the more likely they were to happen! The risk profile is normalized to fit in [0,1] to scale with the EP data.

B.3.3 How to Process Time-Series Data

Not all historical data is so easily obtained in terms of frequencies or statistical samples. Sometimes the data is in the form of a time series, as illustrated by the graphs in Figure B.4 and corresponding Table B.2. Figure B.4a displays the number of people infected by SARS versus time, measured in days from the first reported incident. Time is measured in days, and some days have multiple incidents because they were reported on the same day, but from different countries.

Figure B.4b shows the results of converting the time-series data into frequencies and then into log–log plots to calculate fractal dimension. This conversion is done by the spreadsheet in Table B.2. Bins are shown across the top of

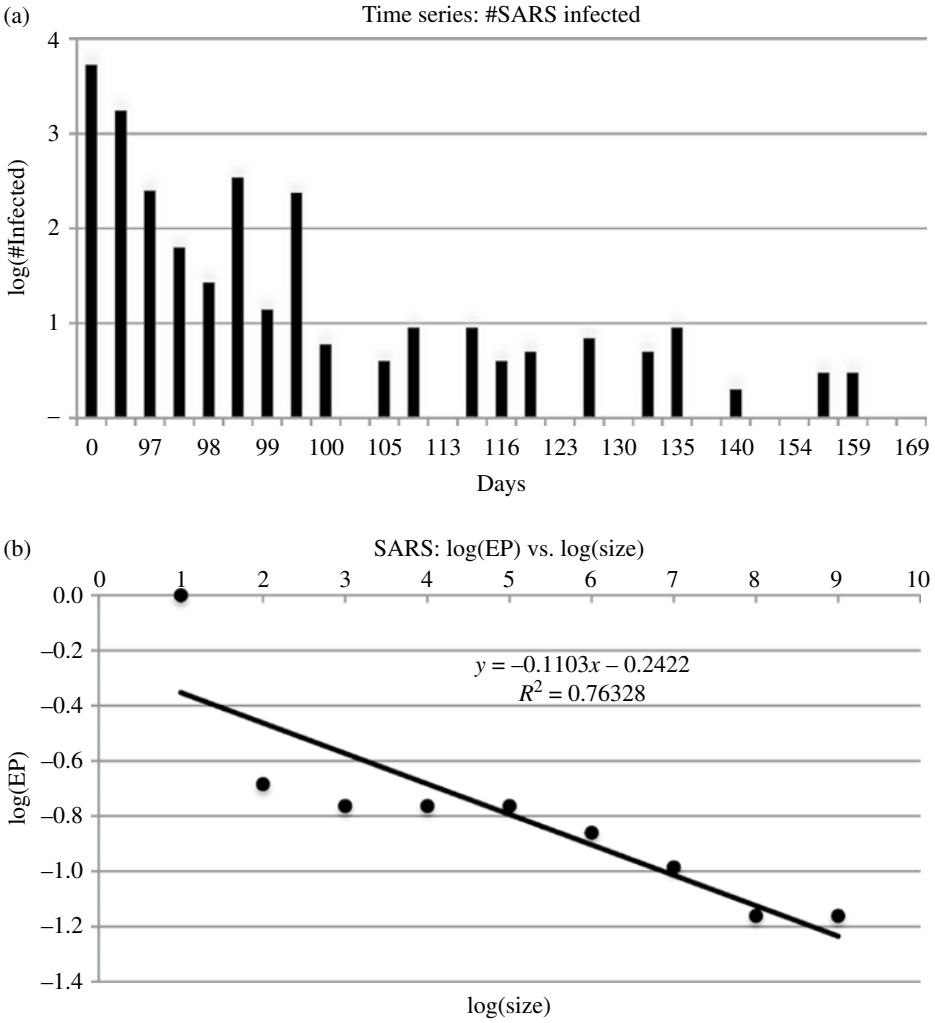


FIGURE B.4 Number of people who contracted SARS during an epidemic that started in China on day zero and spread for 169 days to other countries shows evidence of being a long-tailed hazard, but the data are presented as a time series. It must be converted into a frequency graph and then into an exceedence probability. (a) Time-series data for number of people infected by the SARS outbreak in China and its subsequent spread for the next 169 days. The vertical axis is given in logarithmic scale to make the time series easier to visualize. (b) Logarithmic graph of exceedence probability versus number of people infected indicates a 76% correlation with a power law. Fractal dimension is 0.11, which is very high risk.

TABLE B.2 Spreadsheet containing SARS data and calculations needed to convert time-series data into frequency and exceedence probability data. The fractal dimension is calculated by further conversion of exceedence probability and size into logarithmic values shown at the bottom of the table

Spread of SARS		Counts	50							
Elapsed time (days)	Infected	50	100	150	200	250	300	350	400	450
0	5327	0	0	0	0	0	0	0	0	1
89	1755	0	0	0	0	0	0	0	0	1
97	251	0	0	0	0	0	1	0	0	0
97	63	0	1	0	0	0	0	0	0	0
98	27	1	0	0	0	0	0	0	0	0
99	346	0	0	0	0	0	0	1	0	0
99	14	1	0	0	0	0	0	0	0	0
99	238	0	0	0	0	1	0	0	0	0
100	6	1	0	0	0	0	0	0	0	0
101	1	1	0	0	0	0	0	0	0	0
105	4	1	0	0	0	0	0	0	0	0
113	9	1	0	0	0	0	0	0	0	0
113	1	1	0	0	0	0	0	0	0	0
115	9	1	0	0	0	0	0	0	0	0
116	4	1	0	0	0	0	0	0	0	0
118	5	1	0	0	0	0	0	0	0	0
123	1	1	0	0	0	0	0	0	0	0
125	7	1	0	0	0	0	0	0	0	0
130	1	1	0	0	0	0	0	0	0	0
132	5	1	0	0	0	0	0	0	0	0
135	9	1	0	0	0	0	0	0	0	0
137	1	1	0	0	0	0	0	0	0	0
140	2	1	0	0	0	0	0	0	0	0
143	1	1	0	0	0	0	0	0	0	0
154	1	1	0	0	0	0	0	0	0	0
159	3	1	0	0	0	0	0	0	0	0
159	3	1	0	0	0	0	0	0	0	0
169	1	1	0	0	0	0	0	0	0	0
169	1	1	0	0	0	0	0	0	0	0
	Size	23	1	0	0	1	1	1	0	2
	Frequency	0.79	0.03	—	—	0.03	0.03	0.03	—	0.07
	EP	1.00	0.21	0.17	0.17	0.17	0.14	0.10	0.07	0.07
	log(Size)	1.36	0.00			0.00	0.00	0.00		0.30
	log(EP)	0.00	−0.68	−0.76	−0.76	−0.76	−0.86	−0.99	−1.16	−1.16

the spreadsheet of size 50. There are 9 bins labeled 50, 100, 150, …, 450. How many infected fall into each bin? The 0 or 1 in each column is calculated by the spreadsheet as follows:

$$= \mathrm{IF}\Big(\big(\$\mathrm{B4} < \mathrm{D\$3}\big) * \big(\$\mathrm{B4} >= \mathrm{C\$3}\big), 1, 0\Big)$$

A one is placed in the cell if the number infected in column B falls into the bin. Otherwise, a zero is stored in the cell. To obtain a count of the number of infected people in each bin, sum the columns. This is the Size value shown near the bottom of the spreadsheet. The first bin contains 23 infected reports in the first bin, the second bin contains only 1, and so forth. These counts are converted into a frequency distribution by dividing by the total number of reports.

The exceedence probability row designated EP is calculated as defined by true exceedence probability. Note that it sums to 1.0 in the row designated EP. Now, the logarithms can be taken and displayed in a graph, as shown in Figure B.4b. Excel calculates the slope of this logarithmic plot of values using a regression line and turns out to be 0.11 in this case. However, the R-squared fit is not especially good, because the exceedence probability distribution is not a pure power law. Why?

B.4 NETWORK RISK

The simple PRA approach does not model an interdependent system of components and assets what we define as a *network*. Lewis and Al-Mannai used network theory to model

critical infrastructure such as water, power, energy, transportation, and telecommunications systems as networks. Their model represents a system of assets as an abstract graph, G = {N, E, f}, where N is a set of n assets called nodes, E is a set of m relationships or connections called links, and f: N×E is a mapping of links to node pairs. G defines a network or system of $i = 1, 2, \ldots, n + m$ assets—each with its own threat t_i, vulnerability v_i, and consequence c_i.

Albert et al. [1] studied the vulnerability of structured networks (vs. random networks) in terms of their ability to remain connected. They found that degree sequence $g = \{g_1, g_2, \ldots, g_n\}$ of network structure makes a major difference in the survivability of a network. If the network is random, its degree sequence distribution will obey a *binomial distribution*. In a scale-free network the distribution follows a power law. Scale-free networks can be protected by focusing on high-degreed hub nodes at the expense of less connected nodes. Albert and Barabasi assumed all nodes and links are of equal value, however, which rarely occurs in the real world. Lewis introduced weights, representing consequences, on nodes and links, and extended the Albert–Barabasi model to nodes and links with risk, degree, betweenness, and other network science properties.

Al-Mannai and Lewis [2] give closed-form solutions to the problem of allocating a fixed budget to nodes and links such that risk is minimized, where network risk is defined in terms of a network of *threat–system pairs*:

$$Z = \sum_{i=1}^{n+m} g_i t_i v_i c_i$$

where

g_i = normalized degree of node if asset i is a node and 1 if asset i is a link
t_i = probability of attack
v_i = probability of failure, if attacked
c_i = damage/consequence if asset i fails
n = number of nodes
m = number of links

The normalized degree of a node is computed by dividing each degree value by the maximum degree (hub):

$$g_i = \frac{d_i}{\max\{d_i\}}; \quad d_i = \text{degree of node}_i$$

Al-Mannai and Lewis compared linear and nonlinear models of vulnerability versus allocation and showed that overall network risk is minimized by allocating more resources to nodes and links with higher values of the product: $g_i v_i c_i / v_i(\infty)$, where $v_i(\infty)$ is the cost to eliminate vulnerability of node i. The Al-Mannai–Lewis model considers risk from the defender's point of view and ignores the attacker.

B.5 MODEL-BASED RISK ANALYSIS (MBRA)

MBRA defines network risk Z as the sum of threat–system pair risks weighted by some network property or combination of network properties such as degree and betweenness. Additionally, the MBRA risk model uses simulation to obtain a risk and resilience profile for the entire network. The central model and mathematics for resource allocation are given without proof below. For derivations, see Al-Mannai and Lewis [2].

Network risk is defined differently than fault tree risk, because network risk incorporates a network property such as node degree, betweenness, or height. Node degree is equal to the number of links connecting the node to other nodes. Betweenness is the number of paths passing from all nodes to all other nodes, along shortest paths, going through a node or link. Height is the number of hops from sink nodes to source nodes in a directed network. These properties are normalized by dividing them by the largest value among all nodes and links. For example, node degree is normalized to [0,1] by dividing all node degrees by the maximum degree across all nodes. Similarly, betweenness and height are normalized on [0,1]. Two or more network properties may be combined by multiplying their normalized properties. The combined normalized property g_i, in turn, normalizes risk:

$$z_i = g_i t_i v_i c_i; \quad 0 \le g_i \le 1$$

$$g_i = \prod_{j \in \text{nodes,links}} \frac{w_i(x_j)}{\max\{w_i(x_j)\}}$$

$$w_i(x_j) = \left\langle \begin{array}{l} x_1 : \text{degree} \\ x_2 : \text{betweeness} \\ x_3 : \text{height} \\ \quad \vdots \\ x_k : \text{contagiousness} \end{array} \right\rangle$$

In addition, MBRA employs a modified Stackelberg optimization to allocate attacker resources to increase or decrease threat with the intent to maximize risk and to allocate defender resources to vulnerability and consequence with the intent to minimize risk. This is an extension of the attacker–defender model first proposed by Major [3], Powell [4], and Powers and Shen [5]. The objective of the attacker is to strategically "buy up" threat where it will do the most damage, while the objective of the defender is to "buy down" vulnerability and consequence where it reduces risk as much as possible. MBRA iterates between threat allocation and

prevention and response allocation until reaching equilibrium between threat maximization and risk minimization. (Of course, there is no guarantee of a Nash equilibrium, in which case MBRA stops after several hundred iterations.)

The central model of network risk in MBRA is (see Fig. B.5)

$$Z = \sum_{i}^{n+m} g_i t(a_i) v(p_i) c(r_i)$$

where for each asset (node or link) i

a_i : investment to increase threat

p_i : investment to reduce vulnerability

r_i : investment to reduce consequence

g_i : weight corresponding to one or more user-selected network science parameters

$t(a_i)$: threat function defining the probability of an attack:

$$t_i(a_i) = 1 - \exp(-\alpha_i a_i); \quad \alpha_i = \frac{-\ln(1 - t_i(\infty))}{a_i(\infty)}$$

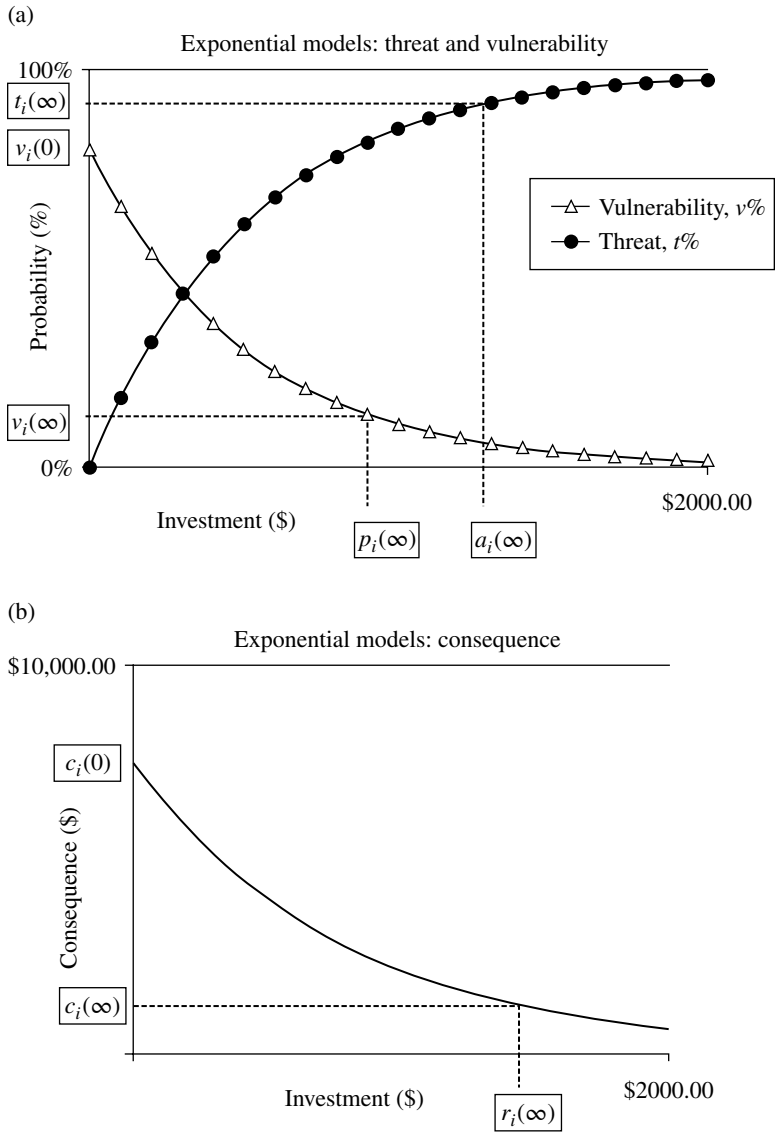


FIGURE B.5 Models of threat, vulnerability, and consequence used by MBRA attempt to represent diminishing returns as exponential functions. (a) Exponential functions used by MBRA to model threat and vulnerability. The slope of each exponential is established by input values shown in boxes. (b) Exponential function used by MBRA to model consequence. The slope is established by input values shown in boxes.

$v(p_i)$: vulnerability function defining the probability of destruction, if attacked:

$$v_i(p_i) = v_i(0)\cdot\exp(-\gamma_i p_i) \quad \gamma_i = \frac{-\ln\left[\frac{v_i(\infty)}{v_i(0)}\right]}{p_i(\infty)}$$

$c(r_i)$: consequence function defining damages from a successful attack:

$$c(r_i) = c_i(0)\cdot\exp(-\beta_i r_i) \quad \beta_i = \frac{-\ln\left[\frac{c_i(\infty)}{c_i(0)}\right]}{c_i(\infty)}$$

n: number of nodes

m: number of links

The network risk function above is constrained by budgets:

$$\mathrm{T} = \sum_{i=1}^{n+m} a_i \quad a_i \geq 0$$

T: attacker's budget, given

$$\mathrm{P} = \sum_{i=1}^{n+m} p_i \quad p_i \geq 0$$

P: defender's prevention budget, given

$$\mathrm{R} = \sum_{i=1}^{n+m} r_i \quad r_i \geq 0$$

R: defender's response budget, given

Figure B.5 shows how inputs are used to calibrate these exponentials using estimates of threat, vulnerability, and consequence for given investments. Functions $t(a_i)$, $v(p_i)$, and $c(r_i)$ mathematically model the increase in t and reduction in v and c, given investments a_i, p_i, and r_i, respectively.

Figure B.6 illustrates a simple network as it appears in the MBRA edit window. Note how an optional map placed in the background facilitates laying out a network over a region of the Earth. This is a useful feature for modeling physical systems such as pipelines, waterways, power lines, roads, and so on. However, the map does not impact the analysis in any way and may be turned off.

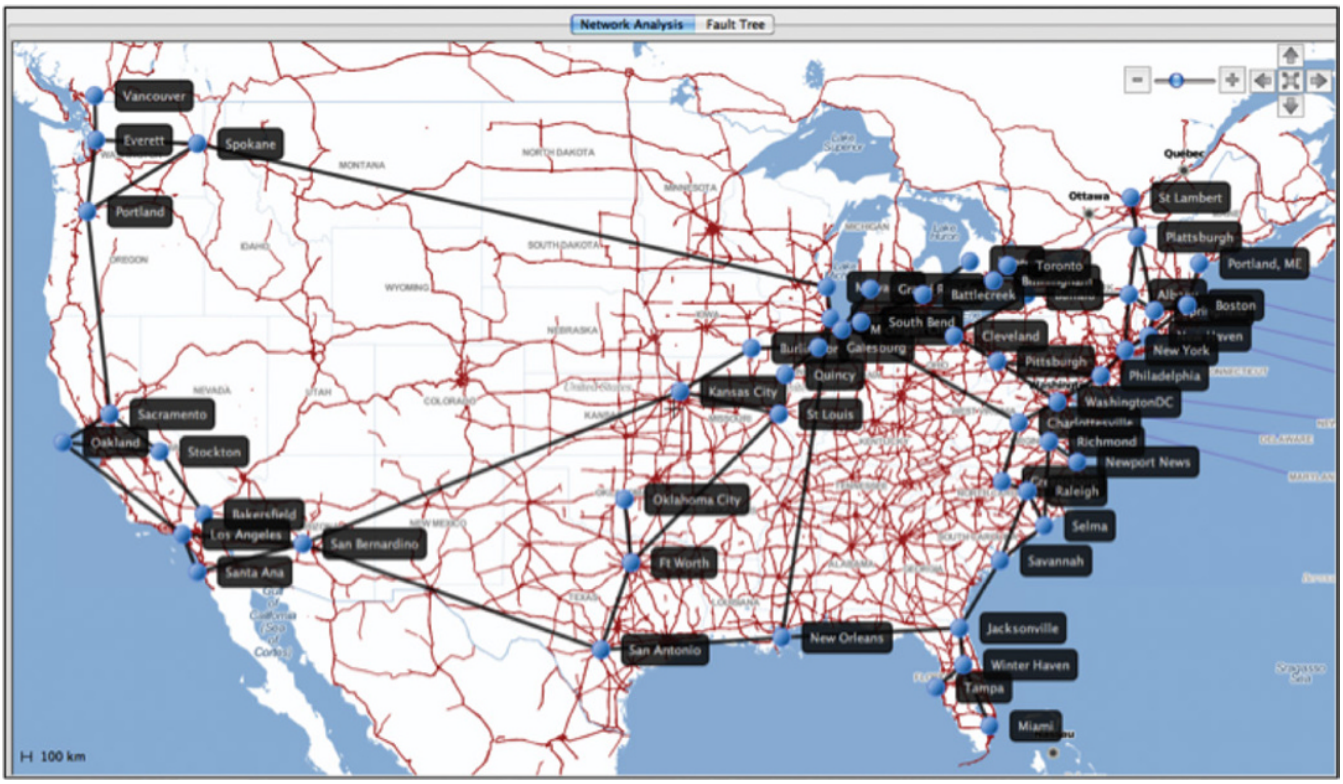


FIGURE B.6 Network model of major Amtrak routes in the United States. Nodes are stations and links are railways.

B.5.1 Network Resource Allocation

MBRA simulates an attacker–defender Stackelberg game where an adversary attempts to maximize network risk and a defender attempts to minimize network risk, Z. Both attacker and defender have a limited and fixed resource to apply toward optimization. The defender's resources are divided into two parts: a prevention budget used to reduce vulnerability and a response budget used to reduce consequence. Threat, vulnerability, and consequence are all modeled as exponential functions as in Figure B.5.

The objective function for resource allocation is

$$\min_{pi,r_i}\left\{\max_{a_i} \mathrm{Z}\right\}$$

Vulnerability and consequence both decline exponentially with investment. So minimization is applied sequentially. On the other hand, the adversary applies the threat budget all at once, knowing the allocation of the defender. The defender observes the allocation of threat against nodes and links of the network and attempts to minimize risk, given the attacker's allocation strategy. This process of allocation, counter-allocation, and reallocation is repeated until a stalemate is reached whereby additional improvement is impossible. This is a Stackelberg game with sequential defender allocations (prevention and response) prior to each attacker allocation.

Stackelberg/Sequential Resource Allocation Algorithm

1.0. Initially Z_0 is risk with no investment, and no iterations.
1.0. Repeat for $t = 1, 2, \ldots$
1.1. Allocate prevention budget P to minimize with respect to vulnerability: $\mathrm{Z}_0 \rightarrow \mathrm{Z}_1$
1.2. Allocate response budget R to minimize with respect to consequence: $\mathrm{Z}_1 \rightarrow \mathrm{Z}_2$
1.3. Allocate attack budget T to maximize with respect to threat: $\mathrm{Z}_2 \rightarrow \mathrm{Z}_3(t)$
1.4. Until there is no change, e.g. $\left|\dfrac{\mathrm{Z}_3(t)-\mathrm{Z}_3(t-1)}{\mathrm{Z}_3(t)}\right| < \varepsilon$

Sequential allocation algorithms are obtained in the same manner as before with fault tree optimization. We use a Lagrange multiplier to remove constraints; differentiate the resulting objective function with respect to prevention, response, and attacker allocations; set to zero; and solve for investment amounts. The resulting closed-form solutions are shown here. Note that each solution is expressed in terms of other solutions. The circular dependency of a solution on previous solutions is why a sequential algorithm is used.

Allocate prevention budget

$$\ln(\lambda_p) = \frac{\sum_{i=1}^{n+m} \dfrac{\ln\left[g_i t(a_i) v_i(0) c(r_i)\right]}{\gamma_i} - \mathrm{P}}{\sum_{i=1}^{n+m} 1/\gamma_i}$$

$$p_i = \frac{\ln\left[g_i t(a_i) v_i(0) c(r_i)\right] - \ln(\lambda_p)}{\gamma_i}$$

Allocate response budget

$$\ln(\lambda_c) = \frac{\sum_{i=1}^{n+m} \dfrac{\ln\left[g_i t(a_i) v(p_i) c_i(0)\right]}{\beta_i} - \mathrm{R}}{\sum_{i=1}^{n+m} 1/\beta_i}$$

$$r_i = \frac{\ln\left[g_i t(a_i) v(p_i) c_i(0)\right] - \ln(\lambda_c)}{\beta_i}$$

Allocate attack budget

$$\ln(\lambda_t) = \frac{\sum_{i=1}^{n+m} \dfrac{\ln\left[g_i v(p_i) c(r_i)\right]}{\alpha_i} - \mathrm{T}}{\sum_{i=1}^{n+m} 1/\alpha_i}$$

$$a_i = \frac{\ln\left[g_i v(p_i) c(r_i)\right] - \ln(\gamma_t)}{\alpha_i}$$

Note that each of the three allocation equations above is expressed in terms of the other two. Prevention allocation $\mathrm{p_i}$ is a function of r_i and a_i, r_i is a function of p_i and a_i, and a_i is a function of p_i and r_i. (If any budget is zero, the corresponding allocation is zero—a calculation that can be made without the allocation equation. However, in the case of threat, a budget of T = 0 means the input values of t_i are used in place of one.) This explains why the Stackelberg optimization is performed sequentially: First, the p_i and r_i are set to zero. The first values of a_i are allocated based on initial p_i and r_i values. Then, new values of p_i are calculated based on previous r_i and a_i. Subsequently, new values of r_i are calculated based on previous a_i and p_i. The process is iterated until there is no change in risk.

Threat, vulnerability, and consequence are driven by budgets. In extreme cases the only inputs needed to perform a complete risk analysis are estimates of consequences and the cost of prevention and response. Rational values of threat and vulnerability can be calculated from such meager input data simply by optimization. That is, assuming a rational actor model, MBRA is able to calculate risk, make optimal allocations, and output expected values of threat and vulnerability (Figure B.7). (Threat and vulnerability are

typically very difficult to estimate. Thus, MBRA addresses one of the major obstacles to risk analysis by calculating t and v values instead of requiring that the user enter them.)

B.5.2 Simulation

MBRA uses simulation in an unusual way—to calculate exceedence probabilities, and in turn, probable maximum loss, PML. Exceedence probability EP(x) is the probability that an event of size x or larger will occur within a system—not just an asset. This is ideal for network analysis because network failures are systemic.

Grossi and Kunreuther [6] explain how exceedence probability is applied to risk assessment using probable maximum loss, PML, in place of expected value. PML risk is the expected loss due to a hazard of size x or larger. It is a worst-case estimate of consequence rather than an average-case estimate.

MBRA uses discrete-event simulation; hence the results are discrete valued. A fault episode is initiated by selecting a single node or link to fail. Then the consequence of the episode is recorded and analyzed. Consequences from k episodes are recorded as a sequence $c_1, c_2, \ldots, c_k$, converted into a frequency histogram, and then normalized into a probability distribution. Finally, the probability distribution is converted into a discrete exceedence probability, EP(c_i). Therefore, MBRA calculates true EP instead of ranked EP:

$$\text{Given}: c_1, c_2, \ldots, c_k$$

$$\text{freq}(c_i): \text{number of times } c_i \text{ occurs}$$

$$\Pr(c_i) = \frac{\text{freq}(c_i)}{\sum_{i=1}^{k} \text{freq}(c_i)}$$

$$\text{EP}(c_{n-i}) = EP(c_{n-i+1}) + \Pr(c_{n-i}); \quad i = 1, 2, \ldots, n-1$$

B.5.3 Cascade Risk

Cascades are commonly occurring fault episodes in many complex systems such as power grids, nuclear power plants, telecommunications systems, and Internet exploits. The idea is simple: Starting with a failed node or link, propagate the fault to adjacent nodes with probability of $t_i v_i$. Sum up the consequences and save in c_j. Repeat episodes until k consequence totals have been recorded. Calculate and display the exceedence probability as shown in Figure B.5.

Each episode starts by selecting a node or link at random and marking it as failed. The selection is random, but proportional to the $t_i v_i$ value of the node/link. This approximates what one would expect: that node/link failure rates should be proportional to likelihood of failure. Similarly, propagation to adjacent nodes is proportional to the $t_i v_i$ value of the adjacent nodes/links.

B.5.4 Flow Risk

Flow risk is very similar; however, it requires a directed network. That is, all links must be unidirectional. Flow simulation simulates the flow of the objective function value through the network from source nodes to sink nodes. A source node is automatically defined as a node with zero incoming links. Sink nodes are defined as nodes with zero outgoing links.

MBRA starts flow simulation by calculating the total values of objective functions across all sink nodes. This is the baseline output of the network. Then MBRA selects one node or link at random (proportional to the $t_i v_i$ value of the node/link) and marks it as failed. This means the node/link will disable all of its outputs to outgoing links. Finally, MBRA recalculates output of the network by simulating

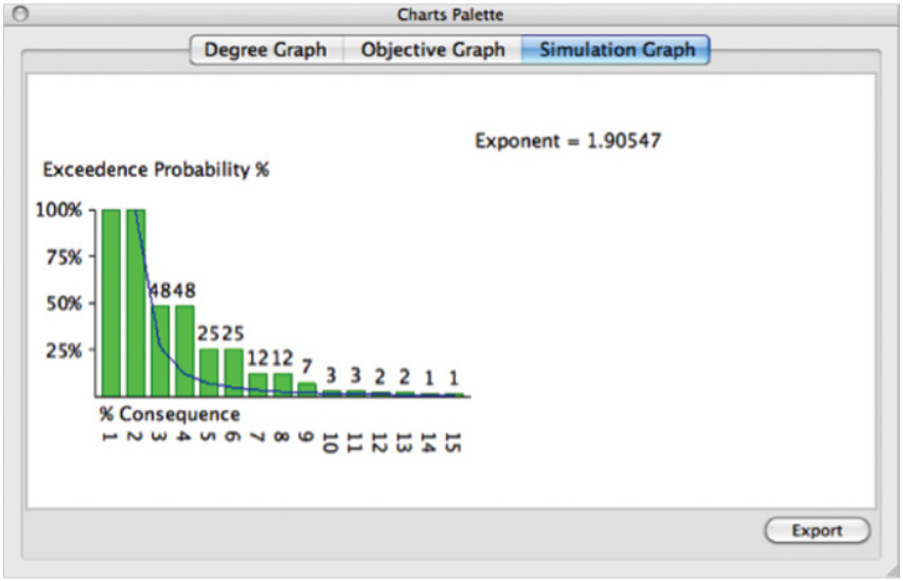


FIGURE B.7 Exceedence probability of cascade episodes is obtained by simulation of the Amtrak network shown in Figure B.6. The exceedence probability fits a power law with exponent $q = 1.9$.

network flow without the marked node or link. This produces a consequence as follows:

$$\text{Given flows}: c_0, c_1, c_2, \ldots, c_k; \quad \text{where } c_0 = \text{nonfailure flow}$$

$$\text{Calculate decreases in total flows}: fc_i = \frac{c_0 - c_i}{c_i}$$

$$\text{freq}(fc_i): \text{number of times } fc_i \text{ occurs}$$

$$\Pr(fc_i) = \frac{\text{freq}(fc_i)}{\sum_{i=1}^{k} \text{freq}(fc_i)}$$

$$EP(fc_{n-i}) = EP(fc_{n-i+1}) + \Pr(fc_{n-i}); \quad i = 1, 2, \ldots, n-1$$

Network flow is simulated by an iterative process whereby the output of each node at time *t* is distributed to all outgoing links proportional to their objective function values set by the user. The proportionality is calculated as the ratio of node objective value versus total output from node *i*:

$$\text{Fraction}_i = \frac{obj_i}{\sum_{j \in \text{outgoing}} obj_j}$$

Inputs from time *t* are used to update outputs at time $(t+1)$. This is repeated until there are no changes in outputs. To perform this computation, MBRA forms a capacity matrix C, from the objective function selected by the user; a fraction matrix F, from the proportionality calculation above; and a state matrix S, defined as the flow through each node.

In matrix form

$$\mathrm{S}_t = \min\left\{\mathrm{C}, \mathrm{F}^{\mathrm{T}} \mathrm{S}_{t-1}\right\}; \quad t = 1, 2, \ldots$$

$$\mathrm{F}^{\mathrm{T}} = \text{transform}(\mathrm{F})$$

$$\mathrm{S} = \begin{bmatrix} \mathrm{s}_1 \\ \mathrm{s}_2 \\ \cdots \\ \mathrm{s}_n \end{bmatrix}$$

$$\mathrm{F} = \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & \cdots & & \\ \cdots & & & \\ & & & f_{n,n} \end{bmatrix}$$

$$\mathrm{C} = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & \cdots & & \\ \cdots & & & \\ & & & c_{n,n} \end{bmatrix}$$

The state equation contains the min{…} function because nodes are not allowed to overflow their maximum capacities. The capacity matrix incorporates the network's topological connections or network structure as well as the maximum values of the objective function selected by the user. Finally, the calculations terminate when there is no longer a change in S, or *n* iterations have occurred.

REFERENCES

[1] Albert, R., Jeong, H., and Barabasi, A.-L. Error and Attack Tolerance of Complex Networks, *Nature*, 406, 378–382, 2000.

[2] Al-Mannai, W. and Lewis, T. Minimizing Network Risk with Application to Critical Infrastructure Protection, *Journal of Information Warfare*, 6, 2, 52–68, 2007.

[3] Major, J. Advanced Techniques for Modelling Terrorism Risk, *Journal of Risk Finance*, 4, 1, 1–9, 2002.

[4] Powell, R. *Defending Strategic Terrorists over the Long Run: A Basic Approach to Resource Allocation*. Institute of Governmental Studies. University of Californian, Berkeley. Paper WP2006-34, 2006.

[5] Powers, M. R. and Shen, Z. *Colonel Blotto in the War on Terror: Implications for Event Frequency*. Fox School Working Paper. Temple University, 2005.

[6] Grossi, P. and Kunreuther, H. *Catastrophe Modeling: A New Approach to Managing Risk*, New York, NY: Springer, 245pp.

APPENDIX C

MATH: SPECTRAL RADIUS

Networks are represented inside of a computer as a matrix. Computer algorithms can then calculate degree, betweenness, cluster coefficient, and spectral radius. (For a more detailed explanation of these algorithms, consult Lewis [1].) This appendix surveys the basics.

C.1 NETWORK AS MATRIX

Nodes and links are represented internally as a square connection matrix C of dimension n, where n is the number of nodes in the network. Figure C.1 illustrates the correspondence between the flow network of Chapter 4 and its connection matrix. As illustrated in Figure C.1, the rows and columns of C can be labeled with the network's nodes for clarity. The elements are either 0 or 1. If a link exists between node n_i and n_j, a one is placed in the cell corresponding to element (i, j). Otherwise, the cell is zero. Note that i and j start at zero, so they are in $[0, n-1]$.

The degree of a node is the sum of its row cells. For example, in Figure C.1, the degrees are simply

$$\text{Degree}(\text{C}) = \begin{matrix} \text{Source} & 1 \\ \text{Intersection} & 3 \\ \text{Bypass A} & 2 \\ \text{Bypass B} & 2 \\ \text{Destination} & 2 \end{matrix}$$

Since nodes cannot be connected to themselves, the diagonals of C are zero. Also, the connection matrix of a bidirectional network is symmetric. The connection matrix of a directional network is asymmetric. Typically, the direction is from left to right in the connection matrix, so source → intersection means a one is placed in the (0,1) cell, but not the (1,0) cell of Figure C.1.

C.2 MATRIX DIAGONALIZATION

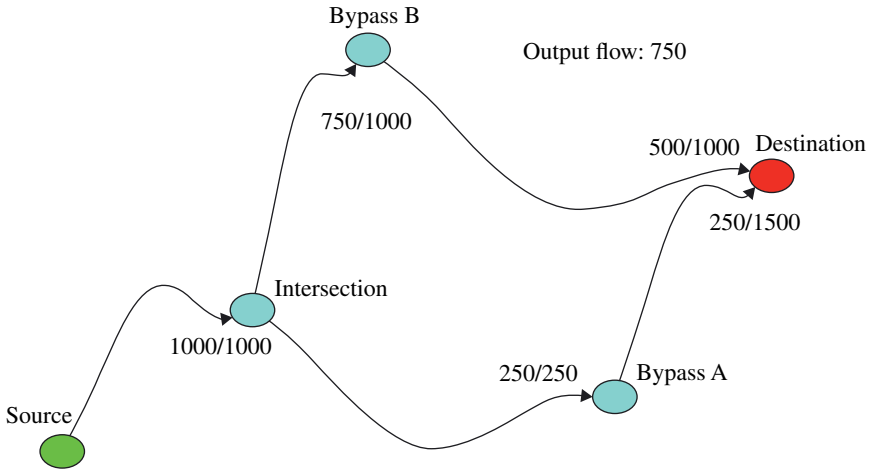
Perhaps the most difficult calculation to perform and understand is the *spectral radius*. In fact, for best results, a computer should always be used to compute the spectral radius of a network. The following description and example is not for the faint of heart. The spectral radius of Figure C.1 will be carried out by hand, but a computer uses a different algorithm than shown here. The reader may want to study matrix algebra and the algorithms for matrix diagonalization.

> ***Spectral Radius:*** *Spectral radius ρ is the largest nonzero eigenvalue of connection matrix C. Eigenvalues lie on the diagonal of a diagonalized connection matrix.*

What is an eigenvalue? Let C be a connection matrix. The spectral radius ρ is obtained by transforming C into a diagonal matrix λI, where I is the identity matrix and λI is the vector containing eigenvalues: λ_0, λ_1, …, λ_{n-1}. That is, we want to find vector λ, such that C = λI. This is done by

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e



Format :

C =	Source	Intersection	Bypass A	Bypass B	Destination
Source	0	⋯			
Intersection	⋯	0	⋯		
Bypass A	⋯	⋯	0	⋯	
Bypass B	⋯	⋯	⋯	0	
Destination	⋯	⋯	⋯	⋯	0

$$C = \begin{matrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{matrix}$$

FIGURE C.1 The flow network of Chapter 4 as it is represented in a computer as a connection matrix (also, see Fig. 4.13).

mathematically solving for the λ vector that makes $\det(C - \lambda I) = 0$. Here we go!

To get started, subtract λ from the elements lying on the diagonal of the connection matrix in Figure C.1, and set its determinant to zero. Reduction of this matrix to its determinant produces a polynomial in λ:

$$\det(C - \lambda I) = \begin{matrix} -\lambda & 1 & 0 & 0 & 0 \\ 1 & -\lambda & 1 & 1 & 0 \\ 0 & 1 & -\lambda & 0 & 1 = 0 \\ 0 & 1 & 0 & -\lambda & 1 \\ 0 & 0 & 1 & 1 & -\lambda \end{matrix}$$

A computer would use a numerical method to find the values of λ that make this determinant equal to zero. The famous *Laplace expansion* by cofactors method is used here. Submatrix A is the cofactor of $C_{0,0}$, and B is the submatrix cofactor of $C_{0,1}$. All other terms in the first row of C are zero, so they can be ignored. Expansion yields

$$\det(C - \lambda I) = (-\lambda)A - B$$

$$A = \begin{matrix} -\lambda & 1 & 1 & 0 \\ 1 & -\lambda & 0 & 1 \\ 1 & 0 & -\lambda & 1 \\ 0 & 1 & 1 & -\lambda \end{matrix}$$

$$B = \begin{matrix} 1 & 1 & 1 & 0 \\ 0 & -\lambda & 0 & 1 \\ 0 & 0 & -\lambda & 1 \\ 0 & 1 & 1 & -\lambda \end{matrix}$$

Repeated applications of Laplace expansions eventually reduce the matrices to polynomials in λ:

$$A = \lambda^4 - 4\lambda^2$$
$$B = 2\lambda - \lambda^3$$

Substitution into the equation $\det(C - \lambda I) = (-\lambda)A - B$ and rearranging terms:

$$\det\left(\mathrm{C}-\lambda\mathrm{I}\right)=\lambda\left(\lambda^4-5\lambda^2+2\right)=0$$

Clearly, one solution to this polynomial equation is $\lambda=0$. There are four others, shown here in rank order:

$$\lambda_0=-2.13$$
$$\lambda_1=-0.66$$
$$\lambda_2=0.00$$
$$\lambda_3=0.66$$
$$\lambda_4=2.13$$

The spectral radius of the network in Figure C.1 is the largest nonzero eigenvalue, which is 2.13 in this case. Note that some eigenvalues are negative. In general, it is possible for all eigenvalues to be negative, in which case the largest eigenvalue is zero. But the definition of spectral radius excludes zero. Hence, spectral radius is the largest nonzero eigenvalue.

What exactly does this mean? First, note that link density matters, because more links means more 1's in the connection matrix, which means a larger spectral radius. Therefore, spectral radius is a measure of *percolation*. Second, the number of 1's in a column or row of C also increases spectral radius. A hub node has more 1's in its row/column than other nodes. Therefore, ρ is also a measure of network structure as defined by "hubness."

An alternative approach to diagonalization involves repeated multiplication of $\mathrm{C}x(t)$ by C, where $x(0)$ is an initial vector approximation to λ. At each stage of this iterative process, $x(t)$ is scaled by α, where α is typically the smallest element of $x(t)$ and $x(t+1)=\alpha\mathrm{C}x(t)$. As t approaches infinity, the scaled matrix, $x(t+1)$ approaches a diagonalized matrix.

The alternative approach is called the *power method* of diagonalization, because it involves raising C to a power. The details of the power method algorithm are not as important as the insight it yields. Each time a connection matrix is multiplied by itself, the resulting matrix represents a hop from a node to its neighbors. C represents one-hop connections; C^2 represents two-hop connections; C^3 represents three-hops; and so on. Therefore, the influence one node has on other nodes is a function of the *reachability* along a chain of hops from one node to others. A node with more reachability will have a larger entry in C^t. The spectral radius can be thought of as overall reachability. More reachability equates with more influence, and influence equates with the impact nodes have on each other in a contagion model of cascades.

Think of spectral radius as an overall measure of the reachability of any node to any other node in a CIKR network. As reachability increases, so does vulnerability to cascades. This is also known of as centrality—degree centrality equates with reachability. The logic behind this statement is that a highly connected node is "more reachable," or "reachable by more nodes," than less connected nodes. In this sense, spectral radius represents the "gravitational pull" of the most central node.

Spectral radius is related to PageRank as used by Google.com to rank order Web pages during a search. The highest-ranking page exerts more influence over lower-ranked pages due to the structure of the network.

C.3 RELATIONSHIP TO RISK AND RESILIENCE

Chapter 4 makes a number of claims regarding the relationship between infectiousness, fractal dimension, and spectral radius. The purpose of this appendix is to add more insight into why these relationships work:

1. $z\sim\gamma\rho$
2. $\log\left(q\right)=b+kz=b+k\gamma\rho$

Equation 1 is the fundamental resilience equation of Chapter 4, and Equation 2 relates resilience to fractal dimension. The y-axis constant b and proportionality constant k depend on the type of network and hazard being modeled. Results in this book assume cascades follow the contagion model of collapse, and all assets are subject to an identical probability of failure, γ. MBRA relaxes this assumption and substitutes TV for γ in cascade failure simulations.

C.3.1 Equation 1

In an ideal network, ρ represents the degree of a typical node. That is, every node is linked to ρ adjacent neighbors. Spectral radius is a proxy for the network's typical node. If a typical node of degree ρ initially fails, then the expected number of neighbors to be "infected" is $\gamma\rho$. Fault spreading increases with $\gamma\rho$ and becomes a certainty on average when $\gamma\rho\geq 1$ but inevitably dies out when $\gamma\rho<1$. This leads to the conclusion

$$z:\left\{\begin{array}{c}<1\ \textit{low}\ \text{risk}\\ \geq 1\ \text{high risk}\\ >>1\ \text{complex catastrophe}\end{array}\right\}$$

Furthermore, the exceedence probability distribution of a certain hazard no longer strictly obeys a power law when $z>1$. For $z\gg 1$, exceedence begins to take on the shape of a normal distribution. Why? The simple answer is that the network becomes saturated when $z\gg 1$, because the likelihood that all nodes will be infected rapidly rises and approaches certainty.

C.3.2 Equation 2

The second fundamental CIKR network equation relates fractal dimension to resiliency. It is an empirically derived equation and, therefore, depends on two constants and the spectral radius of the CIKR network, b, k, and ρ. The constant of proportionality, k, is typically negative. Therefore, $\log(q)$ transitions from positive to negative when $b + kz$ crosses the threshold:

$$b + kz = 0$$
$$z = -\frac{b}{k}$$

This critical point corresponds with the transition from low to high risk, when $z = \gamma\rho > 1.0$. This is called the network's critical vulnerability:

$$\gamma > \frac{-b}{k\rho}; \quad \text{high-risk}$$

For example, using the average values of b and k derived in Chapter 4, $b = 0.5$ and $k = -0.42$, the transition takes place at the critical point:

$$0.5 - 0.42z = 0$$
$$z = \frac{-0.5}{-0.42} = 1.19$$
$$\therefore \gamma\rho = 1.19$$

As a CIKR network transitions from $\gamma\rho = 1.19$ to $\gamma\rho \gg 1.19$, the hazard transitions from low to high and, eventually, to a complex catastrophic state. This makes it possible to categorize CIKR systems as low-, high-, and very-high-risk systems. Now assume the spectral radius and parameters b and k are known for a certain network. Then the critical vulnerability γ_0 is given by the expression

$$\gamma_0 = \frac{-b}{k\rho} = \frac{1.19}{\rho}$$

For example, if $\rho = 4.5$, $\gamma_0 = 26.4\%$. If node and link vulnerability exceed this number, the network is high risk. If vulnerability exceeds this number by a large amount, the network is prone to catastrophic failure.

REFERENCE

[1] Lewis, T. G. *Network Science: Theory and Applications*, Hoboken: John Wiley & Sons, 2009.

APPENDIX D

MATH: TRAGEDY OF THE COMMONS

The tragedy of the commons parable is a metaphor for sustainability or the lack of it. While it is a simple metaphor, its simplicity belies complex behaviors, as examined here. The earliest mathematical models ignored nonlinear paradox of enrichment side effects, which is also developed, here. The original model first proposed by Alfred J. Lotka in 1925 and Vito Volterra in 1926 is presented, first, followed by the modifications made by Crawford Stanley Holling (1930–) and Eberhard Frederich Ferdinand Hopf (1902–1983). The Lotka–Voltera model assumed a linear relationship between predator and prey (cows and grass), while Holling proposed a nonlinear, diminishing returns relationship. Hopf studied the bifurcation that occurs due to a tipping point or critical point in the solution to the nonlinear equations. The Hopf–Holling model is implemented as a Java program—*Commons.jar*—available from the author.

D.1 LOTKA–VOLTERRA MODEL

Let G(t) and C(t) represent the amount of grass and cows available at time t, and let parameters fertilizer_rate, sell_rate, and eat_rate be obvious input parameters. Then, the dynamic response to eating, selling, and fertilizing the commons is given by the Lotka–Volterra predator–prey equations:

$$d\mathrm{G}(t)/dt = \mathrm{G}(t)[\text{fertilize\_rate} - \text{sell\_rate}\mathrm{C}(t)]$$
$$d\mathrm{C}(t)/dt = -\mathrm{C}(t)[\text{sell\_rate} - \text{eat\_rate}\mathrm{G}(t)]$$

This pair of simultaneous differential equations assumes linear relationships as follows. Grass increases linearly with more fertilizer and declines linearly with cattle sell_rate. Similarly, cows increase linearly with eat_rate and decline linearly with sell_rate.

Therefore, the time rate of change of grass is the difference between growth and number of eating cows. The time rate of change of cows is the difference between number of eating cows and sell_rate.

The stability of this system—or lack of it—depends on values of fertilizer_rate, sell_rate, and eat_rate. The predator–prey system can be stable, unstable, or chaotic, depending on these inputs. If the solution is stable, the state space diagram will reach a fixed point and stay there. If the system is unsustainable, zero will be the fixed point. If the system is unstable, it will either oscillate (form an elliptical state space diagram) or blow up (go off of the state space diagram).

D.2 HOPF–HOLLING MODEL

The Hopf–Holling enhancement adds two nonlinearities: (1) a carrying capacity and (2) a nonlinear and diminishing returns response function. The carrying capacity parameter introduces the paradox of enrichment, and the response function introduces a "limits to growth" factor. Essentially, cows eat less, as they get full:

$$d\mathrm{G}(t)/dt = \mathrm{G}(t)\left[\frac{1-\mathrm{G}(t)}{\text{Capacity}}\right] - \mathrm{C}(t)\mathrm{H}(t)$$

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

$$dC(t)/dt = C(t)\left[eat\_rate H(t) - sell\_rate \right]$$

$$H(t) = \frac{G(t)}{1+G(t)}$$

This pair of simultaneous equations says that grass increases according to a logistics growth curve but decreases according to the rate of consumption determined by how much cows eat. But cows respond nonlinearly, $H(t)$, to the amount of grass available.

There is a diminishing returns on how much grass cows can eat. The rate of change in number of cows equals their growth rate minus sell_rate. Growth rate is limited by $H(t)$.

This system has a tipping point that depends on capacity, but also note that rate of change in grass becomes negative if one of the following occur:

$$G(t)\left[\frac{1-G(t)}{\text{Capacity}}\right] < 0;$$

or

$$C(t)H(t) > G(t)\left[\frac{1-G(t)}{\text{Capacity}}\right]$$

The first condition is met when $G(t) > 1$. The second condition is more complicated and depends on capacity. For example, the second condition is met when:

$$\frac{1-G(t)^2}{C(t)} < \text{Capacity}$$

APPENDIX E

MATH: THE DES AND RSA ALGORITHM

E.1 DES ENCRYPTION

Encode:

Use permutation tables to scramble the plaintext to be encoded:

The 56-bit key + tables produce sixteen 48-bit sub-keys: K_1, K_2, …, K_{16}.

Do this 16 times:

Split 64-bit input data into 2 halves, L and R of 32 bits each.

Expand and permute R into 48 bits and XOR with K_i, $i = 1–16$.

Further scramble with a table that produces eight 4-bit blocks.

Permute the result again, then XOR with L and swap L and R.

L and R are joined back together to form the 64-bit pre-output.

Use a table to permute the pre-output one last time.

Decode: Apply sub-keys in reverse order: K_{16}, K_{15}, …, K_1 using the encode algorithm.

Note: XOR is the EXCLUSIVE-OR operation.

E.2 RSA ENCRYPTION

Let a public key P be a pair of integers (n, e) and a private key V be a pair (n, d).

The public and private keys share, $n = p \times q$, where p and q are randomly chosen primes.

Make sure that n is larger than the largest plaintext character you want to encode.

To encrypt a plaintext character m:

Encode $(m) = m^e \ mod \ n$, where *mod* is the modulo function.

To decrypt a ciphertext character c:

Decode$(c) = c^d \ mod \ n$.

How are the numbers n, e, and d in P:(n, e), and V:(n, d) found?

1. Select large prime numbers p and q at random.
2. Calculate the product $n = p \times q$.
3. Choose a number e such that:
   e is less than n, and
   e has no factors in common with either $(p - 1)$ or $(q - 1)$.
4. Find d, such that $e \times d \ mod \ (p - 1) \times (q - 1) = 1$. One way to find d is to search for values of k and d that make this true: $e \times d = 1 + k(p - 1) \times (q - 1)$, for some $k > 1$.

The *mod* operation is simply the remainder of a/b after division. For example, if $a = 8$ and $b = 5$, $a/b = 8/5 = 1$ with a remainder of 3. So, 8 *mod* 5 = 3.

Here is an example. We want to send a secret message containing the date, December 7, 1941—the three plaintext words {12, 7, 41}—from Honolulu to Washington, DC,

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

using $p = 5$ and $q = 11$, $n = 55$, which is large enough to encrypt plaintext words ranging from 0 to 54. Using the algorithm above, we select e less than 55 and make sure it has no factors in common with either $(p - 1) = 4$ or $(q - 1) = 10$. Note that $(p - 1) \times (q - 1) = 4 \times 10 = 40$. The number e must be prime relative to 40. Suppose $e = 3$, which satisfies this requirement (as does many other numbers such as 7 and 11). Because $p \times q = 5 \times 11 = 55$, the public key is P = (55,3), which the sender in Honolulu uses to encrypt plaintext {12, 7, 41} into {23, 13, 6} as follows:

Ciphertext word 1 = 12^3 *mod* 55 = 1728 *mod* 55 = 31 with remainder 23.

Ciphertext word 2 = 7^3 *mod* 55 = 343 *mod* 55 = 6 with remainder 13.

Ciphertext word 3 = 41^3 *mod* 55 = 68921 *mod* 55 = 1253 with remainder 6.

Now we need a private key V = (55, d), where d satisfies the requirement $(e \times d)$ *mod* 40, which is the same as saying $e \times d = 1 + 40 \times k$ for some k. We have already chosen $e = 3$, so we want to find a d and k such that $3 \times d = 1 + 40 \times k$. The smallest value is $d = 27$, for $k = 2$. (Check: $3 \times 27 = 1 + 40 \times 2 = 81$.) Thus, Washington's private key is V = (55,27).Washington, DC, receives the cipher containing code words {23, 13, 6} and uses its private key V = (55, 27) to reverse the encryption, transforming each code word back into plaintext as follows:

Plaintext word 1 = 23^{27} *mod* 55 = 12

Plaintext word 2 = 13^{27} *mod* 55 = 7

Plaintext word 3 = 6^{27} *mod* 55 = 41

Computing large numbers such as 23^{27} can tax even the most capable computer, so we take advantage of the fact that $27 = 3 \times 3 \times 3$ and $23^{27} = ((23^3)^3)^3$. At each step in the calculation, we can apply the *mod* function to reduce the size of the number. Therefore, 23^{27} *mod* 55 = $((23^3)^3)^3$ *mod* 55 = (12,167 *mod* 55)$^3)^3$ *mod* 55 = $((12)^3)^3$ *mod* 55 = (1,728 *mod* 55)^3 *mod* 55 = 12. If we keep reducing the number modulo 55 after each exponentiation, the intermediate result never gets too large.

Note the choice of private key exponent, d, is arbitrary, except that it must be relatively prime to $(p - 1) \times (q - 1)$. We used $d = 27$, but $d = 67$ is also relatively prime to 40, because there are no factors of 67 that are also factors of 40. (Actually, 67 is a prime.) If we used the private key V = (55, 67), we would get the same result: {12, 7, 41}. There are many private keys that decrypt messages produced by P = (55, 3). Does this weaken the cipher?

APPENDIX F

GLOSSARY

The following are selected nontechnical definitions of terms from DHS documents such as the Risk Steering Committee's *DHS Risk Lexicon* of September 2008:

Accidental hazard source of harm or difficulty created by negligence, error, or unintended failure

Adversary individual, group, organization, or government that conducts or has the intent to conduct detrimental activities

Asset person, structure, facility, information, material, or process that has value

Attack method manner and means, including the weapon and delivery method, an adversary may use to cause harm on a target

Capability means to accomplish a mission, function, or objective

Consequence effect of an event, incident, or occurrence

Consequence assessment process of identifying or evaluating the potential or actual effects of an event, incident, or occurrence

Countermeasure action, measure, or device that reduces an identified risk

Deterrent measure that discourages an action or prevents an occurrence by instilling fear, doubt, or anxiety

Economic consequence effect of an incident, event, or occurrence on the value of property or on the production, trade, distribution, or use of income, wealth, or commodities

Hazard natural or man-made source or cause of harm or difficulty

Human consequence effect of an incident, event, or occurrence that results in injury, illness, or loss of life

Incident occurrence, caused by either human action or natural phenomena, that may cause harm and that may require action

Integrated risk management incorporation and coordination of strategy, capability, and governance to enable risk-informed decision-making

Intent determination to achieve an objective

Likelihood estimate of the frequency of an incident or event's occurrence

Model approximation, representation, or idealization of selected aspects of the structure, behavior, operation, or other characteristics of a real-world process, concept, or system

Natural hazard source of harm or difficulty created by a meteorological, environmental, or geological phenomenon or combination of phenomena

Network group of components that share information or interact with each other in order to perform a function

Probabilistic risk assessment a type of quantitative risk assessment that considers possible combinations of occurrences with associated consequences, each with an associated probability or probability distribution

Probability likelihood that is expressed as a number between 0 and 1, where 0 indicates that the occurrence is impossible and 1 indicates definite knowledge that the occurrence has happened or will happen

Psychological consequence effect of an incident, event, or occurrence on the mental or emotional state of individuals

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e

or groups resulting in a change in perception and/or behavior

Qualitative risk assessment methodology set of methods, principles, or rules for assessing risk based on non-numerical categories or levels

Quantitative risk assessment methodology set of methods, principles, or rules for assessing risks based on the use of numbers where the meanings and proportionality of values are maintained inside and outside the context of the assessment

Redundancy additional or alternative systems, subsystems, assets, or processes that maintain a degree of overall functionality in case of loss or failure of another system, subsystem, asset, or process

Residual risk risk that remains after risk management measures have been implemented

Resilience ability to resist, absorb, recover from, or successfully adapt to adversity or a change in conditions

Return on investment calculation of the value of risk reduction measures in the context of the cost of developing and implementing those measures

Risk assessment product or process that collects information and assigns values to risks for the purpose of informing priorities, developing or comparing courses of action, and informing decision-making

Risk assessment methodology set of methods, principles, or rules used to identify and assess risks and to form priorities, develop courses of action, and inform decision-making

Risk management process of identifying, analyzing, assessing, and communicating risk and accepting, avoiding, transferring or controlling it to an acceptable level at an acceptable cost

Risk management cycle sequence of steps that are systematically taken and revisited to manage risk

Risk management methodology set of methods, principles, or rules used to identify, analyze, assess, and communicate risk, and mitigate, accept, or control it to an acceptable level at an acceptable cost

Risk management plan document that identifies risks and specifies the actions that have been chosen to manage those risks

Risk management strategy course of action or actions to be taken in order to manage risks

Risk matrix tool for ranking and displaying components of risk in an array

Risk mitigation application of measure or measures to reduce the likelihood of an unwanted occurrence and/or its consequences

Risk transfer action taken to manage risk that shifts some or all of the risk to another entity, asset, system, network, or geographic area

Risk-informed decision-making determination of a course of action predicated on the assessment of risk, the expected impact of that course of action on that risk, and other relevant factors

Target asset, network, system, or geographic area chosen by an adversary to be impacted by an attack

Threat natural or man-made occurrence, individual, entity, or action that has or indicates the potential to harm life, information, operations, the environment, and/or property

Threat assessment process of identifying or evaluating entities, actions, or occurrences, whether natural or man-made, that have or indicate the potential to harm life, information, operations, and/or property

Vulnerability physical feature or operational attribute that renders an entity open to exploitation or susceptible to a given hazard

Vulnerability assessment process for identifying physical features or operational attributes that render an entity, asset, system, network, or geographic area susceptible or exposed to hazards

INDEX

Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation, Third Edition. Ted G. Lewis.

Companion website: www.wiley.com/go/Lewis/CriticalInfrastructure_3e