5-1-2016

Evidence-based Cybersecurity: Data-driven and Abstract Models

Benjamin Edwards

Benjamin Edwards

Candidate

Computer Science

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Stephanie Forrest, Chairperson

Jedidiah Crandall

Tyler Moore

Steven Hofmeyr

Evidence-based Cybersecurity: Data-driven and Abstract Models

by

Benjamin James Edwards

B.S., Computer Engineering, South Dakota School of Mines and Technology, 2006

B.S., Applied and Computational Mathematics, South Dakota School of Mines and Technology, 2006

DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctorate of Philosophy in
Computer Science

The University of New Mexico

Albuquerque, New Mexico

May, 2016

DEDICATION

To my wife, Lindsey, for her unending support, encouragement, and willingness to tolerate the desert all these years while I worked.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Stephanie Forrest, for her support, patience, and guidance. She has an amazing knack for guiding research towards truly important questions, and for not accepting the easy, incomplete answers to those questions. Without her, my dissertation would have followed the interesting research of others, rather than trying to lead others to what I believe is interesting.

Steve Hofmeyr deserves a great deal of thanks for helping guide many parts of this research to better places. Steve's questions and suggested experiments substantially improved this research, even if they were usually asked a few hours before submission deadlines. Thanks also to Tyler Moore and Jed Crandall for serving on my committee and being involved in my research. Their penetrating questions and guidance helped steer this research to better places.

I would also like to thank my other collaborators. Thanks go to Michel van Eeten and Hadi Asghari for their willingness to share and prepare their painstakingly collected spam data for our analysis, and for giving valuable feedback on that analysis. I would like to thank Robert Axelrod and Alexander Furnas for their expertise and contribution our models of cyberwarfare. Extra thanks go to Robert Axelrod for commenting on and reviewing other chapters before they went to publication.

I would also like to thank the wider security and computer science community for being willing to listen to my ideas and provide feedback. Many conversations in hallways, at the Santa Fe Institute, and at conferences made me rethink my assumptions, what my results meant, and where the research might be taken next.

Special thanks go to the former and current members of Adaptive Computation Lab. Specifically, and in no particular order George Bezerra, Stephen Harding, Michael Groat, Vu Nguyen, Drew Levin, George Stelle, Eric Schulte, Jana Hartman, Cari Martinez, Cynthia Freeman, Teri Oda and David Mohr. They all took valuable time away from their research to give me feedback on rough drafts, practice talks, and crazy ideas. Thanks to them I know what it's like to be part of a research community.

I would like to thank my parents, Bill and Stephanie, my brother Steve, and Sister-in-Law Julie, for their love and support, from pre-school to grad school. They never questioned why I was still in school, and supported all my pursuits.

This dissertation would not have been possible without the unending support of my wife Lindsey, who loved me through my absences, frustration, and grumpiness. Without her I would not have succeeded.

Evidence-based Cybersecurity: Data-driven and Abstract Models

by

Benjamin James Edwards

B.S., Computer Engineering, South Dakota School of Mines and Technology, 2006
B.S., Applied and Computational Mathematics, South Dakota School of Mines and Technology, 2006

Ph. D., Computer Science, University of New Mexico, 2016

Abstract

Achieving computer security requires both rigorous empirical measurement and models to understand cybersecurity phenomena and the effectiveness of defenses and interventions. To address the growing scale of cyber-insecurity, my approach to protecting users employs principled and rigorous measurements and models. In this dissertation, I examine four cybersecurity phenomena. I show that data-driven and abstract modeling can reveal surprising conclusions about longterm, persistent problems, like spam and malware, and growing threats like data-breaches and cyber conflict.

I present two data-driven statistical models and two abstract models. Both of the data-driven models show that the presence of heavy-tailed distributions can make naive analysis of trends and interventions misleading. First, I examine ten years of

publicly reported data breaches and find that there has been no increase in size or frequency. I also find that reported and perceived increases can be explained by the heavy-tailed nature of breaches. In the second data-driven model, I examine a large spam dataset, analyzing spam concentrations across Internet Service Providers. Again, I find that the heavy-tailed nature of spam concentrations complicates analysis. Using appropriate statistical methods, I identify unique risk factors with significant impact on local spam levels. I then use the model to estimate the effect of historical botnet takedowns and find they are frequently ineffective at reducing global spam concentrations, and have highly variable local effects.

Abstract models are an important tool when data are unavailable. Even without data, I evaluate both known and hypothesized interventions used by search providers to protect users from malicious websites. I present a Markov model of malware spread and study the effect of two potential interventions: blacklisting and depreferencing. I find that heavy-tailed traffic distributions obscure the effects of interventions, but with my abstract model, I showed that lowering search rankings is a viable alternative to blacklisting infected pages. Finally, I study how game-theoretic models can help clarify strategic decisions in cyber-conflict. I find that, in some circumstances, improving the attribution ability of adversaries may decrease the likelihood of escalating cyber conflict.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *"When we take action on the basis of an [untested] belief, we destroy the chance to discover whether that belief is appropriate." – Robin M. Hogarth*

Many cybersecurity problems today occur at a global scale, involving nations, corporations, or individuals whose actions have impact around the world. Despite these global, persistent problems, there is limited research on the actual effectiveness of the many interventions that have been proposed or deployed. For example, botnets have been a vehicle for malicious behavior for more than 15 years, but it is unclear whether the most popular intervention, the botnet takedown, has been effective. Most interventions are inspired by deep, hands-on experience with specific attacks and are never evaluated systematically at a large scale. Moreover, as the scope of cyber-insecurity has increased, no one security practitioner can possibly know all of the relevant details associated with the challenges we face today. Thus, there is a need for more explicit and rigorous methods to determine which interventions are effective and which are are not.

Collecting and analyzing large amounts of data has led to new insights in physics,

biology, and economics, but these methods have not been widely applied in security, even as more data on security incidents is becoming available. Collection and analysis of large-scale security data is a crucial component to understanding the global security landscape. Straightforward analysis is rarely possible with security relevant datasets, as substantial work is required to transform unstructured data into meaningful signals. Moreover, relevant measures of security, such as the concentration of infected machines within an organization, are often heavy-tailed (values vary over many orders of magnitude, and whose larger values are not bounded exponentially) making it difficult to separate the effect of interventions from typical random fluctuations [74]. Security data can also be highly dynamic as technologies change. Measurements that were relevant a few years ago, may not reflect current realities. Rigorous data collection, analysis, and modeling are all needed to secure our increasingly interconnected and computationally reliant society.

Despite the growing availability of security data, the data most pertinent to a given security question may not be available. In some cases it simply has not been, or cannot be, collected, or it is privately held. In cases like these it may not be possible to construct data-driven models to study important security questions. In the absence of data, more abstract models can provide insight and reveal fundamental principles. Abstract modeling allows researchers to conduct what-if experiments, which can reveal trends and effects that would be hidden in small-scale, empirical experiments due to the dominance of heavy tails in security data. This type of modeling also allows us to simulate the effect of untested interventions at a low cost. Finally, abstract modeling lets us examine the impact of interventions across many stakeholders and how interventions which benefit one actor may harm others.

This type of rigorous analysis and modeling can yield surprising conclusions. In each of the substantive chapters of this dissertation we[1] present results that often

[1]In this dissertation, we use the plural 'we' as none of this work could have been successfully completed without the collaboration of others. Where appropriate we provide

challenge conventional wisdom. These surprising results demonstrate the need for and power of our approach. Approaches like those described in this dissertation are especially important now when there are more calls for cyber regulation and little understanding of ground truth. It is unclear how initiatives such as the copyright protection provisions in the Trans Pacific Partnership and European data privacy laws will effect the growth of the Internet, individual liberty, and personal security. While we do not tackle these specific questions in this dissertation, our approach provides a framework to begin considering the effect of new Internet policies.

1.1 Contributions and Organization

This dissertation presents new security insights derived from two data-driven models and two abstract models. First, we study ten years of data breaches and find that, despite recent media attention, publicly reported U.S. data breaches have not increased in size or frequency in the last ten years. Next, we analyze a large spam dataset (127 billion messages, sent over ten years), examine risk factors for spam, and quantitatively examine the effect of botnet takedowns and other historical interventions at reducing global spam concentrations. Then, we demonstrate the effectiveness of abstract models by developing a Markov Model of web infections acquired through search engines, and testing hypothetical interventions, suggesting a promising alternative to blacklisting, which has since been incorporated to Google's ranking algorithms. Finally, we develop and analyze game-theoretic models of nation-state cyber-conflict. These models suggest that in the current context it may be rational for the United States to tolerate some cyber attacks. They also suggest that increasing the ability of an adversary's tactical ability to attribute cyber attacks would deter future attacks. Taken together these four models applied to a diverse set of security

footnotes attributing collaborative work.

problems demonstrate the role that careful modeling can play in informing security practices and policy.

Most of the work presented in this dissertation has been presented or published in other venues. Material from chapter 3 was published and presented at the 2015 Workshop on the Economics of Information Security [74]. Chapter 4 was published as "Analyzing and Modeling Longitudinal Security Data: Promise and Pitfalls," which appeared in the *Proceedings of the 2015 Annual Computer Security Applications Conference* [75]. Finally, a version of chapter 4 was published as "Beyond the Blacklist: Modeling Malware Spread and the Effect of Interventions" in the *Proceedings of the 2012 Workshop on New Security Paradigms* [76].

1.1.1 Chapter 2: Background

Chapter 2 gives a general background of the computer security field and its focus on specific attacks and defenses. We show that most security research is found in four areas: identifying new attacks, devising defenses, analysis of current approaches, or developing tools to facilitate research. The review concludes that there is a critical gap in the current research landscape that the work in this dissertation addresses: How can we establish a link between trends in security phenomena and which interventions are likely to or have had a long-term impact on malicious activity?

1.1.2 Chapter 3: Data Breach Hype and Heavy Tails

This chapter studies trends in data breaches over the past decade. Widely publicized data breaches have exposed the personal information of hundreds of millions of people. Some reports point to alarming increases in both the size and frequency of data breaches, spurring institutions around the world to address what appears

to be a worsening situation. We studied a popular public dataset of U.S. breaches maintained by Privacy Rights Clearinghouse and developed rigorous, data-driven statistical models to investigate trends in the dataset over time. We used Bayesian generalized polynomial trend models to investigate the distribution of data breach size and frequency, and we used the models to differentiate between different possible trends. Careful statistical analysis showed that neither size nor frequency of data breaches has increased over the past decade. More importantly, we found that the apparent increases that have attracted so much media attention can be explained by the heavy-tailed statistical distributions that best describe the data. We were able to use our model to predict the probability of major breaches in the future, and we showed that even without increases in breach frequency or size, the heavy-tailed nature of the data indicates that we can expect more large breaches in the future. Finally, we extended the model to include findings from other researchers on the cost of data breaches, and provide estimates of the cumulative cost of data breaches in the next three years.

1.1.3 Chapter 4: Modeling Ten Years of Spam Interventions

Next, we again use data-driven statistical models to study the impact of popular spam interventions such as botnet takedowns. This chapter investigates trends in worldwide email spam from a data set consisting of 127 Billion spam messages sent from 440 million unique IP addresses spread across 260 ISPs in 60 countries over the course of a decade [75]. The data allowed us to identify the concentration of spam sending IP addresses within countries and ISPs. We were then able to use this measure to determine external risk factors for high concentrations and the effect of the numerous interventions designed to fight spam. As with data breaches, we find that spam concentrations are heavy-tailed. This makes determining the relationship between spam concentrations, risk factors, and interventions difficult. Our model

analysis shows that geography, national economics, Internet connectivity and traffic flow all impact local spam concentrations. We then developed statistically robust time-series models, to identify the effect of historical interventions. We used the model to identify three statistically distinct eras within the ten-year data set. We studied twelve different historical botnet takedowns and found that most had little long-term impact on global spam levels, in some cases global spam even increased six weeks after the takedown. Moreover, we found that takedowns have highly localized effects. Takedowns that are highly effective in some countries are followed by increases in other countries at a later date.

1.1.4 Chapter 5: Making Search Safer

Websites are a common vector for the spread of malware. Search engines can play a crucial role in mitigating the spread of malware by directing users away from infected websites. However, falsely identifying an organization's page as infected and removing it from search results erroneously could have serious economic consequences for that organization. Given the need to both protect users and to ensure the appropriate flow of traffic to uninfected websites, how should search providers react to potentially malicious websites? In this chapter, we develop a simple Markov model of malware spread through large populations of websites and study the effect of two interventions that might be deployed by a search provider: blacklisting and depreferencing of infected web pages [77]. The model establishes the effectiveness of each intervention both at reducing user exposure to infected sites (true positives) as well as the traffic that might be lost due to false positives. Once again, we found that when traffic to different sites is heavy tailed, the effect of interventions is difficult to identify. The result is significant because it implies that it will also be difficult to determine empirically whether certain website interventions are effective. However, our model results showed that depreferencing to be an effective alternative to black-

listing as it allows search providers to balance reduction in users exposed to infection and false positives.

1.1.5 Chapter 6: Cyber War and Espionage: The Attribution Problem

In the final substantive chapter, we develop and analyze several game theoretic models of cyberconflict. While some lessons from the cold war and traditional conflict apply to the cyber domain, several factors prevent direct applications of these models. Attribution of cyber attacks to state actors is more difficult because digital evidence is often more complex, malleable and prone to manipulation. Moreover, non-state actors such as crime syndicates, terrorist groups, and patriotic hackers often have capabilities that are comparable to some state actors. Given this uncertain context, how should nation states react to new cyber attacks? In this chapter we develop several models that incorporate the strategic aspects of these unique challenges. The models indicate that in many scenarios it may be rational for countries to tolerate persistent cyber attacks without retaliating, and they suggest that increasing an adversary's technical ability to attribute attacks could reduce the likelihood of future cyberconflict.

1.1.6 Chapter 7: Conclusions

We conclude the dissertation by providing a summary of the results, outlining some opportunities for future work, and providing some final remarks. In particular, we point out some obvious directions in which the presented work could be extended to answer new questions, but we also describe how some long standing questions in cybersecurity might be answered by the approaches in this dissertation. In the

final remarks, we present some of the challenges associated with this approach to security, and urge for increasing the availability of data to researchers and the use of appropriate methods for analyzing data as it becomes available.

Chapter 2

Background: Classification of Security Research

Chapter 1, points to the limited research on the effectiveness of security interventions. To frame the contributions of this dissertation and to illustrate where the gaps are, we organize the different areas of security research into a simple classification consisting of four main categories: Attacks, Defenses, Analysis, and Tools. We further define sub-classifications for defenses and analysis, which, to our knowledge, are the first of their kind [73].

Cyber security has traditionally focused on identifying and rectifying vulnerabilities in the confidentiality, integrity, and availability triad [27]. This traditional focus forms the first two classes of security research: *attacks* and *defenses*. Better understanding of attacks and defenses is undertaken through the third category, *analysis*. Finally, to facilitate research, the development of *tools* is a large and active area of research.

Previous research has categorized vulnerabilities and attacks [27, 122, 35, 249, 190, 113, 142, 314, 125]. Given the relatively large amount of research focused on

categorizing attacks we think adding another categorization is unnecessary. Research in defenses falls naturally into five different subcategories: Detect, Obscure, Isolate, Repair, and Counterattack. We give a more detailed description of each of these categories in 2.3. Further, we surveyed research on new defenses published in four major security conferences in 2014 and 2015 and show in which of these five categories each paper fits. Analysis research is broken into three different subcategories: Measurement, Verification, and Impact, detailed definitions of are given in 2.4.

Examining these data reveals that little work has focused on understanding longterm trends in security phenomena, and the impact of interventions on malicious behavior. Further, the data suggest that most research is focused on developing new attacks and defenses, measuring the prevalence of vulnerabilities or defenses, and formal methods for verifying the security properties of systems.

We take a data-driven approach to validate this classification, by reviewing 288 different abstracts from four major security conferences, classifying each in section 2.1. We give detailed definitions of the classifications, and examples in sections 2.2, 2.3, and 2.4.

2.1 Analysis of one Year of Security Conferences

We validate our classification by analyzing 288 abstracts from the recent meetings of four major security conferences: the Association for Computing Machinery 2014 Conference on Computer and Communications Security (CCS 2014), the 2015 USENIX Security Symposium (USENIX 2015), the Institute of Electrical and Electronics Engineers 2015 Symposium on Security and Privacy (IEEE S&P 2015), and the Applied Computer Security Associates 2015 Annual Computer Security Applications Conference (ACSAC 2015).

These conferences were selected because they have broad scope of topics and are highly rated within the field of computer science [188]. Other highly rated security and privacy conferences such as the International Cryptology Conference and the Computer Security Foundations Symposium are not considered because of their narrower focus.

For each conference, we examine the abstract for each paper presented in the technical portion of the conference. Each abstract is subjectively classified into one of the four main categories: Attack, Defense, Analysis or Tools. In the case of an abstract that is classified into defense or analysis, it is further classified into the subcategories given in sections 2.3 and 2.4. In the case where papers provide multiple contributions, such as developing a new attack and a defense to protect against it, we attempt to identify the primary contribution of the paper. Despite this concise classification, we found no papers that could not be classified almost immediately.

Such a subjective classification is prone to error and individual bias, so, we make the date available and invite others to examine the classification and provide an alternative viewpoint.[1].

The paper classifications can be seen in Tables 2.1, 2.2, and 2.3. We note that while there are 50 papers that fall under the subcategory of measurement, we find only three that contain longitudinal studies. Sosaka et al. examine the evolution of anonymous online marketplaces and speculate on future evolution [254], Leontiadis et al. investigate the evolution of search engine poisoning over four years [166] and Ugarte et al. describe how run-time packers, a tool used by malware writers to obfuscate their code, have changed over time [274]. Understanding longitudinal trends in phenomena is the first step to understanding the effectiveness of interventions, and the sparsity of papers of this type emphasizes a gap in current mainstream security

[1]Data on the classification with titles and specific classifications can be found at `http://cs.unm.edu/~bedwards/data/ConferenceClassification.csv`.

Table 2.1: Categorization of research at four recent major security conferences.

Conference	**Attacks**	**Defenses**	**Analysis**	**Tools**
ACSAC 2015	5	26	12	9
CCS 2014	23	46	25	21
IEEE S&P 2015	11	16	8	20
USENIX 2015	17	19	17	13

Table 2.2: Sub-categorization of analysis research at four major security conferences. * indicates the research was part of this dissertation and is covered in chapter 4

Conference	**Verification**	**Measurement**	**Impact**
ACSAC 2015	0	11	1*
CCS 2014	3	21	1
IEEE S&P 2015	3	4	1
USENIX 2015	14	1	2

research. This gap motivates the work in chapter 3, which is the only rigorous long term analysis of data breaches. Further, though chapter 4 focuses on the impact of interventions against spam, it has a strong measurement component, examining ten years of spam data.

As can be seen, little work focuses on the impact subcategory of analysis. Of the five papers that do, one is by the author of this dissertation (see chapter 4), one is by a collaborator [9] which examines the effectiveness of national botnet cleanup efforts. Of the other three, Liu et al. use machine learning and network activity to predict data breaches [175], Khan et al. examine the impact of typo-squatting on users, and suggest that the impact is minimal [141], and Clark et al. show that there is no evidence that software produced using Rapid Release Cycles does not have significantly more vulnerabilities than software produced on extended release cycles. The fact that 1.7% of the surveyed research is focused on the impact security research highlights the research gap that this dissertation fills

Finally, no papers presented within the last year at major security conferences we analyzed employ abstract models as we do in chapters 5 and 6. In cases where

Table 2.3: Sub-categorization of research on defenses at four recent security conferences.

Conference	Detect	Obscure	Isolate	Replace	Counterattack
ACSAC 2015	2	3	13	8	0
CCS 2014	4	12	7	22	1
IEEE S&P 2015	0	9	2	5	0
USENIX 2015	4	6	4	5	0

data are unavailable, whether because it is in private hands or would be unethical/impossible to collect, these types of models will be crucial to informing future action.

2.2 Attacks

The first class of security research involves devising new attacks against existing systems. This has always been a major focus of security research, and a major focus of certain security conferences such as *Blackhat* and *Defcon*.

The ability to identify and categorize vulnerabilities, attacks, incidents and their consequences has been the subject of extensive research [27, 122, 35, 249, 190, 113, 142, 314, 125]. The resulting taxonomies define a common language for security researchers to analyze current threats, and identify potential new ones. Once classified, new threats may be addressed using similar technical defenses to those already identified in a taxonomy. This has led to a proliferation of taxonomies for specific attacks (Denial of Service [190], Web [314], Intrusion Detection [106]), and a categorization of their various properties [125].

We note that many global cybersecurity concerns, such as data breaches, spam, web malware, and cyber warfare span many layers or hierarchical branches of attack taxonomies. Spam is a canonical example. Spam messages can be sent through

servers with a flawed SMTP policy, compromised mail servers, compromised personal computers, or compromised cloud-based email accounts. Spam messages may attempt to spread malware, solicit money, sell gray market pharmaceuticals, or simply spread misinformation [228]. Each of these vulnerabilities that are exploited to send spam would be classified into a different branch of the attack taxonomy. Also, it is now common to use a variety of methods to infiltrate targets [184]. For example, STUXNET exploited as many as seven different security vulnerabilities including two zero-day exploits [85], spread via removable storage and local area networks, and successfully modified the contents of various systems including personal computers and programmable logic controllers. Rather than attempt to modify these taxonomies (or create our own) to classify these complex problems, we adopt a generic term: **Cyberwickedness**. This term was originally coined by Tyler Moore and introduced in [119]. Cyberwickedness refers to a persistent or chronic cyber threat that may span multiple vulnerabilities, systems, spreading mechanisms, origins, and targets. We will also use this term to refer to the quantity of cyberwickedness originating from or present in a particular entity.

In summary, for the security problems studied in this dissertation, details of the attacks and where they fit in a taxonomy are less important than the fact that these problems have persisted over time and a number of interventions to stop them have been developed.

2.3 Defenses

Developing new defenses for the constant stream of new attacks is another active area of security research. However, in contrast to attacks, to our knowledge, no classification exists that categorizes the types of responses to vulnerabilities and attacks. We assert that most existing corrective cyber security defenses fall into

one of five different categories: Detect, Obscure, Isolate, Repair and Counterattack. We do not suggest that these exactly and exhaustively classify all types of possible security response, but they cover most existing responses of which we are aware, and in our examination of recent security research papers in 2.1, we failed to find an example that did not fit into at least one category. In the following subsections, we illustrate each category with a number of different examples. We use spam as a canonical example of a persistent problem which different types of defenses have been developed to stop.

2.3.1 Detect

Our first class of defense is *detect*. Detect refers to methods for identifying malicious behavior in the system of interest, and it is often the first step in deploying defenses. In some cases detection can be sufficient to protect users by simply alerting them to malicious behavior. Detection is often used in conjunction with other defenses. For example, isolating malicious behavior (see section 2.3.2) often requires first detecting malicious behavior.

Users frequently interact with detection defenses on the web and in email. For example, Google's Safe Browsing is a blacklist service that detects phishing and malware serving websites, and warns users before they are allowed to access potentially malicious web pages [104]. The first step in spam filtering is detecting malicious emails, and in many cases spam emails are allowed to appear in a user's inbox with a warning about spam rather than the email itself being isolated in a spam folder [61]. Other systems have been developed to detect the circumvention of international calling rates via Voice over IP links [230], identifying malicious accounts in online communities [261], and detecting when Internet of Things devices are tampered with using Channel State information [15].

2.3.2 Isolate

Isolate is a defense in which a computer or subnetwork is isolated from other systems to prevent malicious agents from communicating and/or compromising the system. Isolation of systems can come in many forms, from simply shutting down vulnerable systems [258], to complex sandboxing mechanisms that allow full functionality of components with careful restrictions on the type and amount of communication between different systems [271]. Some types of isolation requires the separation of malicious behavior from legitimate behavior. In this case detection is a prerequisite, and false positives may be a concern [251]. We study the effect of false positives in chapter 5.

A simple form of isolation is to quarantine a compromised system. This may involve disabling a vulnerable system from all use, both legitimate and malicious, preventing damage to other systems. This form of isolation is often performed on critical computing infrastructure through the complete disconnection of the system from any form of network communications. This practice is known as *air gapping* [244]. Air gapping was used to isolate computers at an Iranian nuclear facility, however this was overcome through the use of the STUXNET malware which was transmitted through infected USB drives [86]. Another example of quarantine is in response to exploits in the Java programming language which allow an attacker to seize control of a host computer, Oracle suggested disabling any Java plugins currently installed in a browser [215, 158]. One draconian response to spam is disabling an email address that is emitting spam [61], or blocking an IP address or domain known to send spam [75]. In the case of confidentiality leaks, forcing servers to remove the information may be the only practical course of action, for example, when it was discovered that personal information could be recovered from publicly released data on movie preferences, a court order forced sites hosting the data to remove it [203]. In 2011, Verisign suggested to the International Corporation for Assigned Names

and Numbers(ICANN) that Verisign should have the power to cancel or transfer the registration of domains exhibiting malicious behavior [197]. Censoring is a form of isolation. For example the Chinese microblogging site Weibo, removes posts it deems sensitive from the site [311]. Isolation has been studied as a mechanism for combating network worms [299, 57].

A more subtle version of isolation involves adding barriers while maintaining functionality. The most common realization of this type of isolation is sandboxing of executing programs. In many cases, communication across existing processes can lead to compromise. This is especially true when applications are all run by a single virtual machine as in the case of the Android operating system [13] or the execution of javascript in a browser [271].

When loss of functionality is unacceptable, and sandboxing is not a viable solution, new strategies must be employed. In this case, it is sometimes sufficient for a third party to filter the results of an exploited vulnerability from some users. Spam filters are an excellent example, where despite the large volume of spam messages being sent (tens of billions daily [228]), most end users rarely see spam in their email, and are often alerted to its provenance. Removing potentially dangerous websites from search results is another example of filtering malicious activity. A website hosting malicious software can still be accessed directly, but most users are spared exposure to a potential attack if the sites are removed (filtered) from search results [77]. This type of response is frequently referred to as blacklisting.

Detection is a prerequisite for the filtering type of isolation, and distinguishing between malicious and normal behavior can be difficult. Any system that attempts to filter malicious behavior must consider the trade off between false positives and false negatives. For example, an overzealous spam filter may result in many false positives, blocking potentially legitimate traffic and reducing email functionality. Conversely, a conservative spam filter might allow too many spam messages through, reducing

its usefulness. Balancing these trade offs is key to effective filtering, and this makes filtering a particularly attractive area in which to employ a graduated response, in which the degree of filtering is deployed proportionally to the certainty that a message or activity is malicious [77]. Graduated responses, sometimes referred to as rate limiting, have been effective at reducing the impact of distributed denial of service attacks [129, 302], stopping network intrusions through delaying system calls [252], and resisting malware propagation [293, 296].

2.3.3 Obscure

When appropriate barriers cannot be erected to prevent attackers from accessing confidential information, another approach is to *obscure* relevant information. The main realization of this defense is encryption, though there are other approaches as well.

Encryption continues to be an active area of research, with several major conferences devoted to its development, and we won't attempt to given an entire overview of the field here. d'Agapeyeff provides an excellent introduction to the history of the field [65], Schneier provides an authoritative textbook on the subject [243], and current research on encryption is the topic of two major conferences, the International Cryptology Conference and Annual International Conference on the Theory and Applications of Cryptographic Techniques..

Other methods of obscuring information exist however. Including, address space layout randomization [247], and other fine grained automated diversity techniques such as instruction set randomization [17]. The former obscures the location of critical portions of memory by randomizing their location and preventing attackers from having reliable information about their location. The latter, instruction set randomization, creates a unique and private instruction set for each executing programming,

preventing attackers from designing binary code injection attacks.

More coarse-grained diversity such as N-version programming is also a defense which obscures vulnerabilities by producing N versions of a program, not all of which would contain the vulnerability [39]. In its original form, N-version programming called for the generation of N functionally equivalent programs from the same specification by different programmers. More recent works suggest that N-version programming could be achieved automatically using an evolutionary approach to software [245]. Other work has focused on automatically creating and executing N versions of complete operating systems using address and instruction space randomization and running each in parallel [60]. This defense has been extended beyond software to the design of integrated circuits [4].

Non-cryptographic privacy preserving methods such as negative surveys and k-anonymity hide obscure information while preserving critical properties that allow for later analysis. Negative surveys work by having individuals report information which does *not* describe themselves [84]. With few assumptions, aggregate statistics about the data can be computed without individuals ever having to divulge potentially sensitive information. K-anonymity is a technique that prevents attackers from using multiple pieces of data to identify previously anonymous data by suppressing portions of the data so that individuals cannot be distinguished from groups of size k [164]. Other work on non-cryptographic methods of obscuring information include steganography (hiding data within other data such that both are recoverable) [34], obfuscation techniques to conceal the purpose or author of code [248], and introducing errors into measurements to obfuscate location information [6]. These non-cryptographic privacy preserving methods are the focus of several conferences such as the *Privacy Enhancing Technologies Symposium* and the *Workshop on Information Hiding and Media Security.*

2.3.4 Replace

The most obvious and prevalent response to a security vulnerability is *replacing*. This is the process of replacing a vulnerable mechanism in a system with a version that does not contain that vulnerability. While simple to state, the process of identification, repair and deployment can be difficult. Indeed, much of computer security is focused on the technical challenge of doing just this. After identifying the exploit in Java mentioned above, once a repair was found, Oracle recommended that the software be updated, restoring its functionality [156]. Similarly, an SMTP server running vulnerable software may become compromised and used to send spam messages. If the server is to stay in service the compromised software must be patched [171]. Unfortunately, this form of defense often leaves systems vulnerable for long periods of time until patches can be found. One study found that it takes, on average, 45 days to develop patches for new vulnerabilities [8].

Beyond simple patches to existing software, replacement can also take the form of substituting whole components for more secure ones. Most protocols that serve content over the internet, can be replaced with 'secure' versions in which the data is exchanged after a Transport Layer Security (TLS) handshake takes place. For example, the common email protocol POP3 sends information in plain text, but can be replaced with a version that uses TLS, and establishes an encrypted channel before transmitting data [304]. However, the TLS protocol and its many implementations have been the source of many bugs, and replacements for the protocol itself are frequently suggested [161, 134]. Replacement can also serve to optimize previously secure, but unusable components, encouraging more widespread use [87, 288].

Replacement can occur at multiple scales, from patching software, replacing components with those that have identical interfaces, or at a larger scale replacing entire systems with others that have similar but different functionality. In this case, a new

system is found that replicates some or all of the functionality of the previous system but has fewer security flaws. For example, as a response to the number of vulnerabilities in the Internet Explorer browser, it is often recommended to replace it with a more secure browser such as Firefox or Chrome [20]. Often it is not just software that needs to be replaced, but communication protocols. Because the original SMTP protocol did not authenticate senders it could easily be exploited by spammers, and alternatives to the protocol implementing a variety of authorization techniques have been suggested as replacements [250, 43]. In a multi-user Linux environment, NIS authentication can be replaced with LDAP/Kerboros for user authentication. This is recommended because NIS allows a single compromised user access to all password hashes [220]. Replacing a single system with a number of diverse semantically equivalent variants is another strategy for securing a system [205].

2.3.5 Counterattack

A final category of response, *counterattack*, does not address any particular technical problem. Rather, it focuses on creating a response that actively fights current infections, or punishes those who exploit vulnerabilities. This is the common approach to dealing with botnets [109] and is accomplished by compromising the attackers' computers (referred to as the command and control nodes) that are actively exploiting unsuspecting users' machines. This does not require users to patch the vulnerabilities that initially allowed the attacker to control the host, but it renders the botnet non-functional. The persistence of the original vulnerability is an obvious drawback of this approach. In the case of spam, one counter attack is legal pressure, which has been applied to notorious spammers leading to prison sentences and hefty fines [305]. DNS takedowns are another form of counterattacks, and work by redirecting traffic destined for a specific URL to an IP address controlled by counterattackers. It can be used to prevent access to illegal gambling sites, sites hosting content that violates

copyright, or sites selling illegal products [100].

Counterattacks in many cases are a legal gray area. While they may be effective at disabling or misleading attackers, they do so at the risk of violating current laws, leaving counter attackers open to the same type of prosecution as the attackers. Because this type of response can be as varied as the original attacks, modeling it in a general way is difficult. While further research on the efficacy of counter attacks is warranted, for these two reasons we do not investigate it further in this work.

2.4 Analysis

The third category of security research is *analysis*. We use the broad term analysis to refer to security research that focuses on the exploration of the properties of attacks, defenses, where they arise and their relationship. Analysis includes three categories Verification, Measurement, and Impact. We argue that there is a vibrant community working in verification, and a growing amount of research focused on measurement, but little work studying the impact of attacks, defenses and their relationship.

2.4.1 Verification

Verification refers to mathematically rigorous analysis of security protocols or implementations of software to prove that protocols or implementations are free of vulnerabilities. This subfield can be traced back to a 1986 essay by Donald I. Good calling for the development of logical foundations for computer security [101], and the subsequent founding of the Computer Security Foundations Symposium.

Verification work has large scope, encompassing the more general problem of verifying software systems. Even before Good's essay, work on formal verification

of software can be traced back to work by Pnuelli on assuring the correct operation of programs [219]. Further work by Emerson and Clarke developed model checking, showing that program properties could be expressed as temporal logic formulas and used to verify the execution of parallel programs [81]. This approach was applied to develop services which were robust against byzantine (arbitrary or malicious) faults [242].

Formal verification of hardware [160] and software [23], is an active area of research. Verification of security techniques has existed for more than 30 years, for example verification of the RSA encryption protocol was performed as early as 1984 [31]. More recent work has proven the security of Secure Shell's (SSH) implementation of the signed Diffie-Hellman cypersuite [24], and high performance implementations of elliptic curve cryptographic software [40].

Automated provers such as Coq [185] and EasyCrypt [21] have been crucial to this endeavor, and they have been used to verify OpenSSL hash functions [25] and implementations of private information retrieval protocols [260].

2.4.2 Measurement

The next subcategory of analysis is *measurement*. This type of research studies the distribution of vulnerabilities and defenses across time or systems. It can also include studies of the usability of different security constructs and the relative performance of different security systems. Measurement studies could serve as motivation for other types of research, for example, in a study of top-ranked domains, a measurement study revealed that nearly a third were susceptible to cross-site scripting attacks [165]. This discovery might lead to new defenses to isolate malicious scripts or tools to automatically deploy known defenses.

Measurement studies of defenses can focus on how well defenses are deployed

across different systems. For example, A study of implementations of OAuth in 4,000 Android apps found that 86.2% of apps mis-use OAuth, leading to to potential vulnerabilities [286]. Measurement research can focus on two aspects of defenses at once. For example, a recent study of pattern-based pins (passwords implemented as a series of connected lines) for the Android operating systems measured both their strength and usability [11].

Studies such as the ones cited above, while important, often lack a long-term view of security phenomena. There have been some recent longitudinal studies such as work investigating the evolution of online anonymous marketplaces like the Silk Road [254] and obfuscation tools used to hinder attempts to reverse engineering malware [274, 68]. To understand the impact of different defenses on attacks, one must first understand how the distribution of attacks and defenses is changing over time, including when interventions might be deployed. Without this baseline, it will be impossible to differentiate typical changes from those that are the result of interventions. Longitudinal studies like the ones above and the ones we describe in chapters 3 and 4 provide important clues to understanding the impact of interventions.

2.4.3 Impact

The final category, and one that we argue has been underrepresented in the security community is *impact*. Research in this category focuses on understanding the relationship between deployed defenses, the real-world outcomes of attacks, and any associated externalities. And, in the end, this is what matters when trying to protect users from the malicious activities of attackers.

Focus on impact is relatively rare in security research presented at mainstream conferences, as we demonstrate in section 2.1. Some recent exceptions include a study of effectiveness of national botnet cleanup initiatives on the Conficker botnet [9].

Other work in a similar vein found that Rapid Release Cycles for Firefox are not more prone to vulnerabilities than more traditional software release schedules [46]. Prediction focused research uses network traffic properties in concert with machine learning techniques to predict breaches at different organizations [175].

Research that explores the impact of attacks and defenses has often been published in specialized security publication venues, e.g. The Workshop on the Economics of Information Security [163, 162].

2.5 Tools

The final subcategory in security research is *Tools*. This crucial area of research develops new tools aiding in the research into other categories or the application of defenses for security practitioners. Tools can help identify new avenues for attacks, such as frameworks for symbolic execution of C/C++ code which can identify bugs [227]. Similarly, tools can be used to automatically deploy defenses, for example Script Inspector is a tool for website administrators that checks and isolates third-party scripts [310]. Tools like Snort can help system adminstrators set up their own rules for detecting malicious activity [233].

Tools often aid in measurement and verification as well. EasyCrypt is a proof assistant designed specifically to help prove the security of implementations of cryptographic protocols [21]. Other tools such as Autocog are a tool which measures how accurately android app permissions are described versus how they are actually used [223].

2.6 Summary

This chapter has provided an overview of the security field and how the contributions of this dissertation fill an important gap in current security research. To illustrate this gap we presented a categorization of security research into four categories: Attacks, Defenses, Analysis, and Tools. We argue that two of the three subcategories of Analysis, impact and longterm measurement, have been largely ignored in the top tier general security conferences. This has impeded understanding of the impact of various defenses on malicious behavior. The work in this dissertation will fill this gap through the use of data-driven and abstract models to study longterm trends in malicious behavior and the impact of interventions.

Chapter 3

Hype and Heavy Tails: A Closer Look at Data Breaches[1]

3.1 Introduction

In February 2015, the second largest health insurer in the United States, Anthem Inc., was attacked, and 80 million records containing personal information were stolen [183]. A few months later, the US Office of Personnel Management announced that personal information, including the background checks of 21.5 million federal employees was compromised [307]. Ten months earlier, in September 2014, Home Depot's corporate network was penetrated and over 56 million credit card numbers were acquired [38, 159]. These incidents made national headlines, the latest in a string of large-scale data breaches ([62, 150, 90]) that have spurred both the United States Congress [59] and the White House [145] to propose new disclosure laws to address what appears to be a worsening situation.

[1]The substance and much of the writing in chapter appeared in 2015 Workshop on the Economics of Information Security. I conceived and executed the research for this paper, with my co-authors advising and helping with the final written presentation.

Several studies provide evidence that the problem of electronic data theft is growing. A recent report by TrendMicro concludes that the frequency of data breaches has increased since 2009 [124]. A report published the same month by Gemalto, indicates that the total number of breaches increased by 10% while the number of records breached in the first half of 2015 declined compared to 2014 [96]. A 2014 Symantec report noted that there was an increase in the number of large data breaches, and a dramatic five-fold increase in the number of identities exposed over a single year [55]. In another study, Redspin reported that the number of breaches in the health care industry increased 29% from 2011 to 2012, and the total number of records compromised increased 138% for 2012 to 2013 [128].

But, is the problem actually growing worse? Or if it is, how much worse is it, and what are the trends? As we asserted in chapters 1 and 2, we need rigorous measurements to understand trends in security phenomena. The data used to produce these kinds of reports have very high variance, so simply reporting average values, as in these earlier reports, can be misleading. Figure 3.1 plots breach sizes over the past ten years using data obtained from a popular dataset published by the Privacy Rights Clearinghouse (PRC) [51]. In the figure, data breach sizes span eight orders of magnitude, which means that the average value can be significantly affected by just a few data points. For example, if we consider the identical data, but plot it on a yearly basis, it appears that breaches have increased in average size since 2013 (blue line on the figure). However, this trend is not at all obvious if we consider the data on a monthly or even quarterly basis, also shown in Figure 3.1 (green and red lines). Thus, there is a need for statistically sound data analyses to determine what, if any, trends exist, and where possible to make predictions about the future.

To address these issues, we adopt a statistical modeling approach and apply it to the PRC data, showing that in this dataset neither the size nor the frequency of breaches has increased over time. We use a Bayesian approach, which allows us to

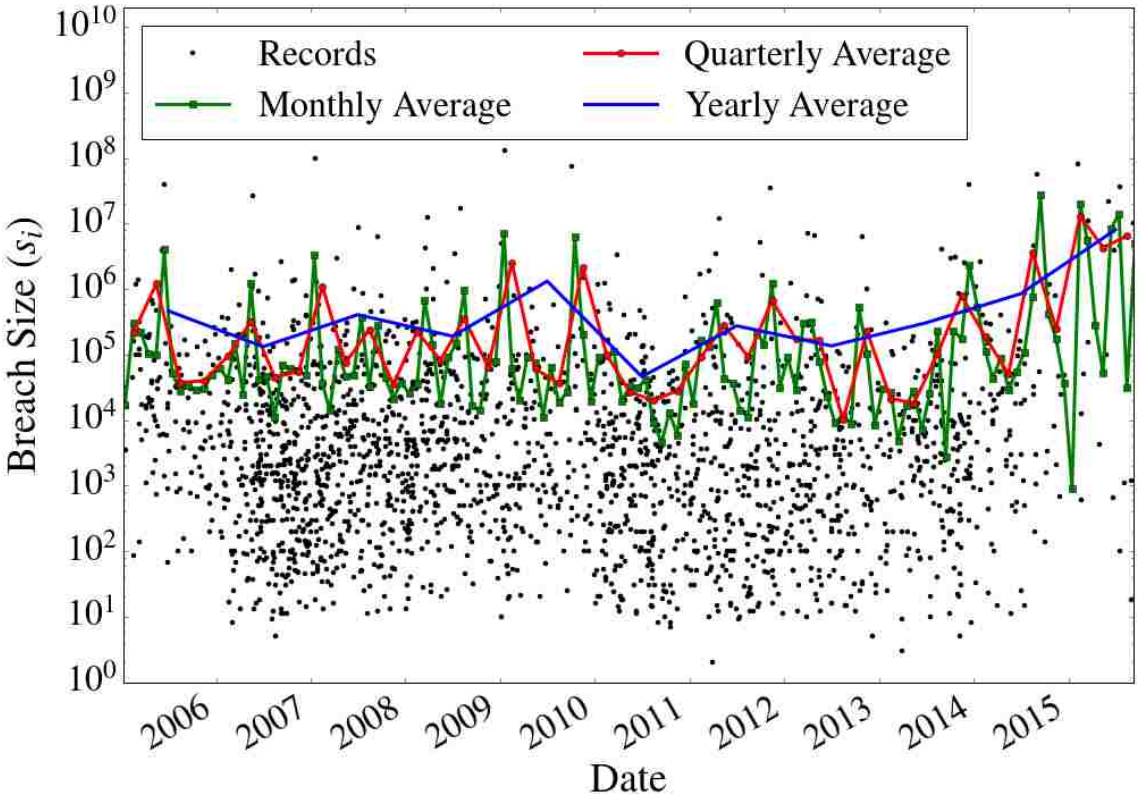

Figure 3.1: Data breach sizes (records exposed) over a ten-year period. Data taken from [51]

construct accurate models without overfitting (see subsection 3.3.1). Our analysis shows different trends for different subsets of the data. We consider two distinct types of breaches: *malicious*, where attackers actively target personal information, and *negligent*, which occur when private information is exposed accidentally (e.g. misplacing a laptop). In the dataset, both the size of malicious and negligent breaches have remained constant over the last ten years. Similarly, the frequency has also remained constant (see subsection 3.3.2 and subsection 3.3.3).

Beyond assessing trends, this approach enables us to determine the likelihood of certain future events, at least in the United States (see section 3.4). For example, the model predicts that in the next three years there is 25.7% chance of another

Anthem sized (80 million) breach, and only a 4.0% chance of a Anthem and Home Depot sized breach occurring within a year of each other. However, there is an 75.6% chance of a breach of at least five million records in the next year. The probabilities are relatively high for breaches of five million records because the distributions that best describe the size of breaches in the dataset are heavy-tailed, meaning that rare events are much more likely to occur than would be expected for normal or exponential distributions.

Another contribution of this chapter is identifying the particular forms of the underlying distributions, which may offer insight into the generative processes that lead to data breaches. For malicious breach sizes, we find that the distribution is log-normal (see subsection 3.2.2); such distributions are known to emerge from multiplicative growth. In fact, the size distribution of companies is best described by a log-normal [264], so we speculate that as a company grows, the number of data records it holds grows proportionally, and breach sizes follow along. We find that negligent breaches are better described by a log-skewnormal distribution [114]. The log-skewnormal distribution is similar to log-normal distribution except it allows for a further skew of the data towards larger breaches. This skew may represent different underlying features of breaches at different organizations. By contrast, the breach frequency for both negligent and malicious breaches best fits a negative binomial, which could be generated by a mixture of different types of breaches, with each type occurring at a different but constant rate (see subsection 3.2.3).

Some of our results seem counter-intuitive given the current level of concern about privacy and the damage that a data breach can cause. However, some simple anecdotal observations about our data lend credence to the results. The largest data breach in our data occurred back in 2009 when cyber-criminals stole 130 million credit card numbers from Heartland payment systems [210].

We used the publicly available dataset that we believe is the most complete,

but our models could easily be applied to additional datasets, for example, datasets that are not yet in the public domain or those that may arise if new disclosure laws are passed. Moreover, by establishing a baseline, the models we describe could be extended in the future by incorporating additional data on the nature of the breaches, which could help identify promising areas for technical improvement. Such analysis could also help policy makers make better decisions about which problems are most pressing and how they should be addressed. For example, cybersecurity today is often framed in terms of risk analysis and management [212, 28]. Accurately assessing risk, however, requires quantitative measures of likelihood and cost. In this chapter, we use available data and statistically sound models to provide precise estimates of the likelihood of data breaches. Using these estimates, we then incorporate two different cost models (see subsection 3.4.4) to assess likely future risks. Depending on the cost model, if trends continue we can expect the cumulative cost of data breaches to be between $4 and $179 billion over the next three years.

3.2 Data

In this section, we describe the dataset obtained from the *Privacy Rights Clearinghouse* (PRC) and examine the distribution of breach sizes and frequencies. We show that the size distribution is well-fit by a log-normal or log-skewnormal distributions, whereas the daily frequency of breaches is well-fit by a negative binomial. Finally, we show how these distributions are affected when the data are divided into malicious and negligent breaches.

3.2.1 Privacy Rights Clearinghouse

The PRC is a California nonprofit corporation focused on issues of privacy [50]. The PRC has compiled a "Chronology of Data Breaches" dataset[2] that, as of September 15, 2015, contains information on 4,571 publicized data breaches that have occurred in the United States since 2005. For each breach, the dataset contains a number of variables including: the date the breach was made public, the name of the entity responsible for the data, the type of entity breached, a classification of the type of breach, the total number of records breached, the location (city and state) where the entity operates, information on the source of the data, and a short description of the breach.

Of the 4,571 breaches in the dataset, only those involving exposure of sensitive information have associated record counts. We restricted our analysis to this subset, which consists of 2,253 breaches. There are two noteworthy limitations to these data. First, the number of records listed in the dataset for each breach is only an estimate of the number of individuals affected, and second, the dataset contains only those breaches that have been publicly acknowledged. However, the PRC dataset is the largest and most extensive public dataset of its type. It is possible that many data breaches are going unreported. Different surveys have indicated that anywhere between 60% [270] to 89% [44] of security incidents go unreported. However, these reports are based on informal surveys of security professionals, their accuracy can't be confirmed (section 3.6), and there is no obvious reason why their size/frequency distributions should differ from PRC.

[2]Available for public download from `http://www.privacyrights.org/data-breach`.

3.2.2 Breach Size

We denote the distribution of breach sizes over the number of records contained in individual breaches as S. For each individual breach i, we denote the number of associated records as s_i. To determine the time-independent distribution that best fits the data, we examined over 20 different distributions, for example, log-normal, log-skewnormal, power-law, generalized pareto, log-logistic, and log-gamma.[3] In each case, we estimated the best fit parameters for the distribution using the maximum likelihood, and then performed a Kolomogorov-Smirnov (KS) test to determine if the parameterized distribution and the data were statistically significantly different [182]. Figure 3.2 shows the fit to log-normal; the KS test gives $p = 0.21$, which means that we cannot reject the null hypothesis that the data were generated by this distribution.[4] For all other tested distributions, $p < 0.05$, which tells us that the data were unlikely to have been generated from that distribution. Although the best fit is to the log-normal, we can see in Figure 3.2 that the data points in the tail (high values of records) deviate from the best-fit line. We return to this issue in section 3.6.

Log-normal distributions often arise from multiplicative growth processes, where an entity's growth is expressed as a percentage of its current size, independent of its actual size [191]. Under this assumption and at steady state, the distribution of entity sizes is known to be log-normally distributed. For example, this process has been used to model the size distribution of companies as measured by annual sales, current employment, or total assets [264]. We speculate that a related process is operating here, if the number of sensitive (customer) records held by a company is proportional to its size, or the number of stored records is increasingly multiplicatively over time.

[3]Specifically, we tested all of the distributions in the scipy stats package that have a domain defined for values > 0. `http://docs.scipy.org/doc/scipy/reference/stats.html#continuous-distributions`.

[4]In this case, higher values of p are better, because they indicate that we are *not* rejecting the null hypothesis, i.e. that the data are drawn from a log-normal.

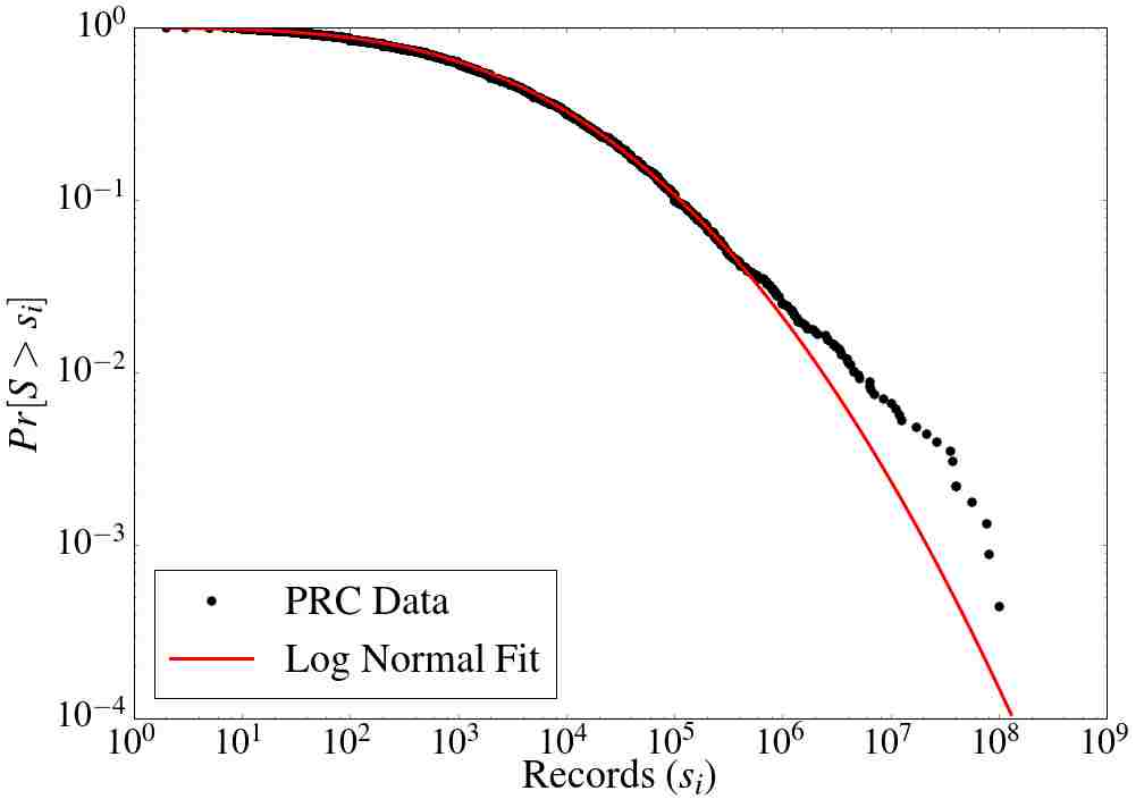

Figure 3.2: The distribution of breach sizes and the fit to a log-normal distribution.

3.2.3 Breach Frequency

We are interested in studying how often breaches occur and whether or not there are interesting trends in breach frequency. The dataset reports the exact date at which each breach became publicly known. For the majority of dates in the dataset, however, there were no publicly reported data breaches, and on days when breaches did occur, there were seldom more than two (Figure 3.3).

We used a similar approach to the one we employed in subsection 3.2.2, except that we studied discrete distributions, because the range of daily frequencies is so small. We examined a number of discrete distributions, such as Poisson, binomial,

zero-inflated Poisson and negative binomial, and found that the best fit is provided by a negative binomial. Figure 3.3 shows that the parameterized negative binomial and the data do not differ significantly, according to the KS test for discrete distributions [7], with $p = 0.99$. If we assume that breaches occur independently and at a constant rate, then we would expect the daily frequency to be a Poisson distribution [111]. However, the data are more dispersed than can be explained by a Poisson, which has a very poor fit, with $p = 8 \times 10^{-10}$.

There are a number of random processes that generate a negative binomial distribution [309]. The most likely candidate in this case is a continuous mixture of Poisson distributions, which occurs when events are generated by a Poisson process whose rate is itself a random variable. In our case, breaches at different organizations, perpetrated by different groups could all have different rates, leading to the negative binomial distribution we observe here. It is also possible that breaches are announced on specific dates to reduce their]impact in the media. This could lead to a clustering of breach reports on Fridays or before holidays.

3.2.4 Negligent and Malicious Breaches

Each breach in the PRC dataset is categorized into one of seven different categories (plus the category *Unknown*). The seven categories naturally divide into two groups. The first are breaches arising from *negligence*, where records were not actively sought by an attacker but were exposed accidentally, for example, through the loss of laptops, or accidental public exposure of sensitive information. The second group includes breaches arising from *malicious* activities that actively targeted private information, for example, attackers hacking into systems, an insider using information for malicious purposes, or payment card fraud. Table 3.1 contains information on the number of each type of breach in the dataset, and our groupings. It is apparent that negligent

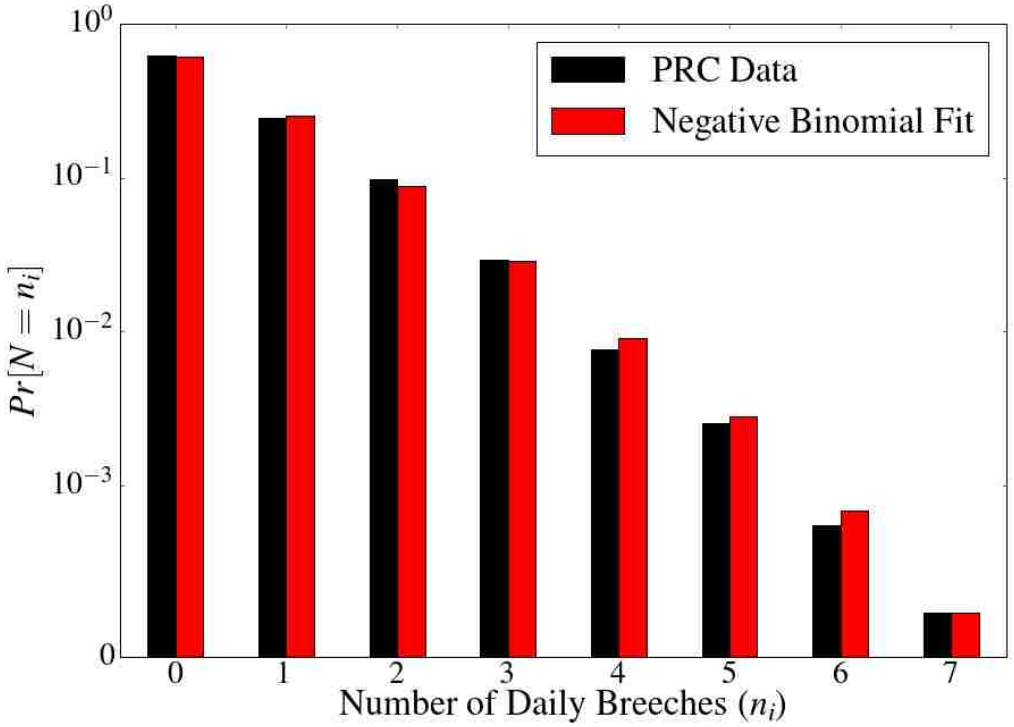

Figure 3.3: The distribution of the daily number of breaches and the fit to a negative binomial.

breaches occur nearly twice as often as malicious breaches.

We re-applied the data fitting analysis described earlier (subsection 3.2.2 and subsection 3.2.3) separately to each of the two groups. We find that even when the data are divided into negligent and malicious categories, each category matches a negative binomial distribution for daily frequency, although with different means. However, malicious and negligent breaches fit different distributions. Specifically, the sizes of malicious breaches are well fit by a log-normal distribution, while negligent breaches are well fit by a log-skewnormal distribution. Even though the lumped data (all categories aggregated) are log-normally distributed, it is possible that the different distributions arise because this distribution is changing over time, or that

different processes are producing different breach sizes. We provide evidence against the former hypothesis in the next section.

3.3 Modeling Data Breach Trends

Our earlier analysis does not allow for the possibility that the distributions are changing over time. In this section, we describe how we use Bayesian Generalized Linear Models (BLGMs) [95] to construct models of trends in the PRC the dataset. We then use Bayesian Information Criteria (BIC) to determine the highest likelihood model, while avoiding overfitting. We use the distributions derived in section 3.2, as the basis for our time-dependent models.

3.3.1 Bayesian Approach

We illustrate our approach by focusing on the sizes of negligent data breaches, S_n. The basic strategy assumes an underlying type of distribution for the data (e.g., sizes of negligent breaches), which we found to be log-skewnormal in subsection 3.2.2. Hence $S_n \sim LogSkewNormal(\mu, \tau, \alpha)$, where μ is the location parameter, τ is the shape parameter (the inverse of the variance), and α is the skew parameter.

To incorporate temporal variations, we model the location parameter, μ, as a polynomial function of time, t, i.e. $\mu = \beta_0 + \beta_1 t + \cdots + \beta_d t^d$. Time is expressed as a decimal value in years since January 1, 2005, with a resolution of one day, e.g. $t = 1.2$ would be March 13, 2006. We describe how to determine the degree of the polynomial, d, later. The parameters, β_i, for the polynomial, together with the shape parameter and skew parameter (τ and α respectively), comprise the free variables of the model. For each free parameter we need to define a prior distribution.

Table 3.1: Types of data breaches as categorized by the PRC, grouped into negligent and malicious breaches.

Breach Type	**Description**	**Count**
Negligent Breaches		1412
Portable Device	Lost, discarded or stolen, portable device or media	627
Unintended Disclosure	Information posted in a publicly available place, mishandled, or sent to the wrong party	456
Physical	Lost, discarded, or stolen non-electronic records	196
Stationary Device	Lost, discarded or stolen stationary device or media	135
Malicious Breaches		781
Hacking	Electronic entry by an outside party	469
Insider	Someone with legitimate access intentionally breaches information	282
Payment Card Fraud	Fraud involving debit and credit cards that is not accomplished via hacking	30
Unknown	Other or Unknown	58

The choice of prior distributions is an important and active area of research in Bayesian statistics. As suggested in the literature [95], we used normally distributed priors for the polynomial parameters, $\beta_0 \sim \mathcal{N}(\overline{\log(S_n)}, 1)$ and $\beta_i \sim \mathcal{N}(0, \frac{1}{Var[t^i]})$, a gamma-distributed prior for the shape parameter, $\tau \sim Gamma(1, 1)$, and a generalized student's T distribution for the skew parameter, $\alpha \sim T(2.5, 0, \frac{pi^2}{4})$ [22]. These priors are "uninformative," i.e. they assume the least amount of information about the data. Although there are other possible priors, our results did not vary significantly when tested with other reasonable choices. Once the model is defined, we can numerically determine the parameters using maximum-likelihood estimation.

To assess the accuracy of the estimates, we determine confidence intervals for the values of the parameters using a variant of Markov Chain Monte Carlo (MCMC) sampling to ensure robust, fast samples [120]. MCMC is an efficient general method for sampling possible values for the parameters of the model.

The remaining unknown in the model is d, the degree of the polynomial. We determine a model for each $d \in [0, 6]$, and choose the model (and hence the polynomial) with the minimum Bayesian Information Criterion (BIC) [246]. We compute the BIC as $BIC = -2L + k * log(n)$, where L is the likelihood of the model when the parameters are set to their MLE, k is the number of parameters (the degree of the polynomial plus any shape parameters), and n is the number of data points. The BIC balances the likelihood of the model, which is increased by adding parameters, with the number of parameters and size of data, and hence prevents overfitting. This enables us to chose a model that best fits changes in the data, rather than modeling statistical noise. This is an important feature when the distributions are heavy-tailed. Another common model selection tool is Akaike Information Criteria (AIC), but we obtained the same results using AIC.

To summarize, our modeling approach involves the following steps:

1. Define a BGLM similar to Equation 3.1, as shown in subsection 3.3.2.

2. Find the maximum likelihood estimates for the parameters of the model (e.g. β_i, τ) for polynomial trends d up to degree 10.

3. Select the model that has the minimum BIC for the maximum likelihood estimates of the parameters.

4. Sample from the distribution of free parameters (i.e. β_i, τ, α) using MCMC to determine the confidence intervals for the parameters.

5. Randomly sample the model to generate a distribution, and compare that to the actual distribution, using the KS test.

3.3.2 Modeling Breach Size

As derived in subsection 3.3.1, the model for negligent breach sizes is

$$\begin{aligned} S_n &\sim LogSkewNormal(\mu, \tau, \alpha) \\ \mu &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_d t^d \\ \beta_0 &\sim \mathcal{N}(\overline{\log(S_n)}, 1) \\ \beta_i &\sim \mathcal{N}(0, \frac{1}{Var[t^i]}) \\ \tau &\sim Gamma(1, 1) \\ \alpha &\sim T(2.5, 0, \frac{\pi^2}{4}) \end{aligned} \tag{3.1}$$

For malicious breaches we fit a similar model, except using a log-normal distribution

$$
\begin{aligned}
S_n &\sim LogNormal(\mu, \tau) \\
\mu &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_d t^d \\
\beta_0 &\sim \mathcal{N}(\overline{\log(S_n)}, 1) \\
\beta_i &\sim \mathcal{N}(0, \frac{1}{Var[t^i]}) \\
\tau &\sim Gamma(1, 1)
\end{aligned}
\tag{3.2}
$$

The best fit model for both malicious and negligent breaches, as determined by the minimum BIC, gives $d = 0$, which indicates that the distribution of sizes is constant. Figure 3.4 shows the median values for models, plotted against the PRC data[5]. Maximum likelihood estimates for the parameters are given in Table 3.2.

Table 3.2: Maximum likelihood estimates and 95% confidence intervals for models of breach size.

Variable	**Estimate**	**95% Confidence Interval**
Negligent		
β_0	6.186	[5.453, 8.111]
τ	0.098	[0.075, 0.139]
α	0.959	[-0.11, 1.521]
Malicious		
β_0	8.052	[7.827, 8.282]
τ	0.093	[0.084, 0.103]

To summarize, we find that the distribution of negligent and malicious breach sizes has remained constant with a median size of 383 and 3141 respectively over the ten-year period represented by the dataset. Random samples generated using Equation 3.1 and the estimates found in Table 3.2, indicate that the predicted distribution of sizes by the model does not significantly differ from the data, i.e. our model generates data that are indistinguishable from the actual data. The KS test

[5]We show median rather than the mean because it better represents the typical values in heavy tailed distributions.

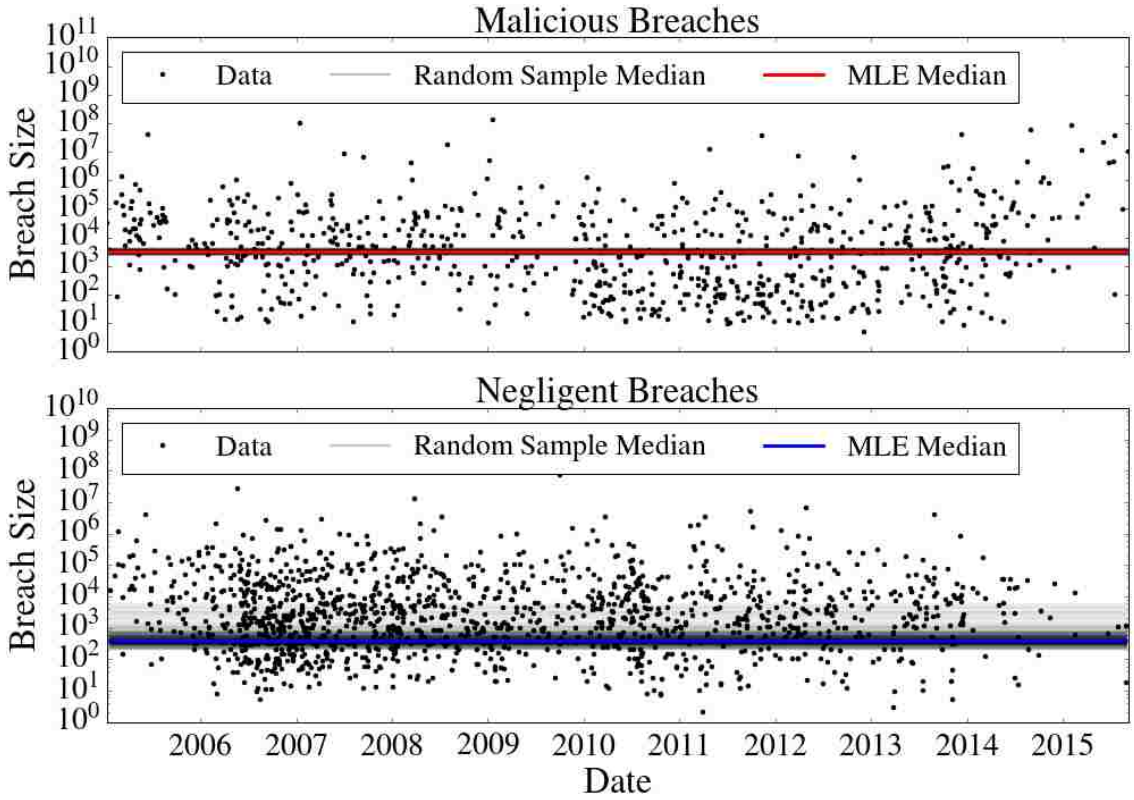

Figure 3.4: The size of data breaches from the PRC dataset, versus the maximum likelihood estimate of the median size.

gives $p = 0.55$ for the fit to the negligent breach sizes, and $p = 0.11$ for the fit to the malicious breach sizes.

3.3.3 Modeling Breach Frequency

We use the same methodology to model the frequency of data breaches, with a negative binomial as the basic distribution, as determined in subsection 3.2.3.[6] The daily frequency, B_n of negligent breaches is given by

[6]We also test a Poisson model, but found it had a higher BIC than a negative binomial model.

Table 3.3: Maximum likelihood estimates and 95% confidence intervals for models of daily breach counts. We report e^{β_0} as this is the mean number of breaches of each type per day.

Variable	**Estimate**	**95% Confidence Interval**
Negligent		
e^{β_0}	0.364	[0.343, 0.388]
α	0.944	[0.762, 1.170]
Malicious		
e^{β_0}	0.200	[0.191, 0.216]
α	1.897	[1.157, 3.107]

$$
\begin{aligned}
B_n &\sim NegativeBinomial(\mu, \alpha) \\
\log(\mu) &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k \\
\beta_0 &\sim \mathcal{N}(\overline{\log(D_n)}, 1) \\
\beta_i &\sim \mathcal{N}(0, Var[t^i]) \\
\alpha &\sim Gamma(1, 1)
\end{aligned}
\tag{3.3}
$$

The same model is used for malicious breaches, replacing B_n with B_m, the daily number of malicious breaches. We use a log link function for the mean value of the negative binomial distribution, which ensures that the mean value is always positive regardless of the value of the polynomial [95].

For the daily frequencies of both negligent and malicious breaches, the models with the lowest BIC are polynomials of degree $d = 0$, indicating that the daily frequency of breaches has remained constant over the past ten years. The maximum likelihood estimates and 95% confidence intervals are shown in Table 3.3. Random samples generated using the Equation 3.3 are not statistically significantly different from the data for both negligent and malicious breaches; which have $p = 1.0$ and $p = 0.99$, respectively, for the KS test.

3.3.4 Modeling Large Breaches

It is possible that the models developed above are dominated by smaller breaches, which have experienced little change over the last ten years, while larger breaches are increasing in size or frequency. We define *large* breaches as those involving 500,000 or more records. This threshold was chosen because it includes a large enough sample size for us to fit reasonable models (93 malicious and 121 negligent breaches), but the threshold is high enough that the breach would likely be reported widely in the press.

Using this definition, we find that large breach sizes fit a log-normal distribution for both negligent and malicious breaches, and that large breaches both categories do not show a significant trend over the past ten years.

The frequency of large breaches, both malicious and negligent, fits a Poisson distribution, rather than the negative binomial observed for breaches of all sizes. This could indicate that different processes are responsible for generating large versus small breaches. Alternatively, it could simply be that the very low probability of a large breach results in a distribution that is difficult to distinguish from the negative binomial. In this case, we would expect the BIC of the Poisson model to be lower because it has one less parameter than the negative binomial. Regardless of whether the best model mathematically is a negative binomial or Poisson, the trends for large breaches are the same as the overall trends, with the frequency of malicious and negligent large breaches remaining constant over the ten years covered by the dataset.

3.4 Prediction

The power of a good statistical model is that it can be used to estimate the likelihood of future events. In this section we discuss what types of predictions models like ours can legitimately make, and point out some of the ways in which naive interpretations of the data can lead to erroneous conclusions. We then demonstrate how the model can be used to quantify the likelihood of some of the large breaches that were experienced in 2014, and we make some predictions about the likelihood of large breaches in the future. Finally, we project the possible cost of data breaches over the next three years.

3.4.1 Variance and Prediction

Because the distributions of both the breach sizes and frequencies in the PRC dataset are heavy-tailed, it is difficult for any model to make precise predictions about the exact number of breaches or their average size. This is different from a dataset that is, for example, normally distributed, where, with sufficiently large sample size, one can say with high probability that samples in the future will cluster around the mean, and estimate the chances of samples falling outside one standard deviation from the mean. However, in the PRC dataset, common statistics like the mean or the total number of records exposed are much less predictable. The data often vary wildly from year to year, even if the process generating the breaches has not changed at all. This phenomenon is common in many complex systems, including many security-relevant datasets, e.g., [76].

We illustrate the effect of the high variability in Figure 3.5 and Figure 3.6. These figures show the result of measuring the total number of malicious and negligent breaches and the total number of records contained in those breaches annually for the historical data (black line) and a single simulation using the models presented

in section 3.3 (red line)[7]. Although our model indicates no trend in the size or frequency of breaches, the distribution can generate large year-to-year variations. These changes are often reported as though they are significant, but our results suggest that they are likely artifacts of the heavy-tailed nature of the data.

For example, a number of industry reports, some using the PRC dataset, have pointed to large changes in the size or number of data breaches from year to year [280, 55]. One of the most alarmist is the Symantec Threat Report which noted a 493% increase in the total number of records exposed from 2012 to 2013, and a 62% increase in the number of breaches in the same time frame.[8] The 493% number includes the large Court Ventures data breach, which was initially reported as revealing 200 million records, but later reports reduced that that number to 3.1 million records [90]. Even with this correction, the report implies a 282% increase in the total number of breached records. These increases sound startling, and a naive interpretation might suggest that both the number and size of data breaches are skyrocketing.

We can test for the likelihood of such extreme changes using our model. To do so, we used the model to generate 10,000 samples of possible annual totals, both for the number of breaches and the number of records, from 2005-2014. We find that a 62% year-to-year increase in the total number of breaches is relatively common in simulation, occurring 14.0% of the time. Similarly, an increase of 282% in total records occurs in 17.6% of year-to-year transitions. These results suggest that the large changes identified in these reports are not necessarily significant and could be natural variations arising from the underlying observed distributions of data breaches.

Although our model cannot accurately predict the total number or typical size of data breaches in any given year, it can assess the likelihood of different sizes of

[7]We use data through 2014 as it was the last complete year we have data

[8]These reports use a combination of public and private data, so comparison of exact numbers is not feasible.

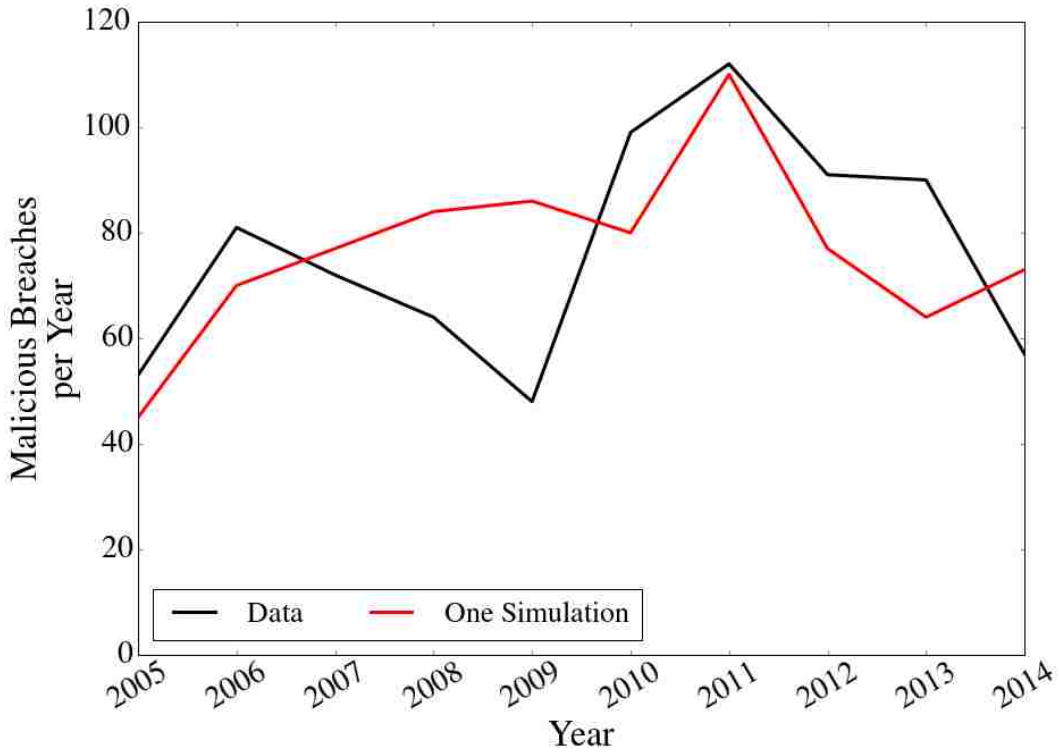

Figure 3.5: The number of malicious breaches reported each year throughout the dataset, together with a single simulation sampled from our model.

breaches. That is, we can predict the probability of a breach of a specific size within a given time-frame, as we show in the next subsection.

3.4.2 "Predicting" the Last Year of Breaches

To assess the likelihood of the breaches that occurred in 2014, we fit the model using data from 2005 to the September of 2014, and used it to "predict" the events of the last year. The MLEs of this smaller dataset are virtually identical to those found for the whole range, suggesting that the 2014 data are not significantly different from those of the previous nine and a half years.

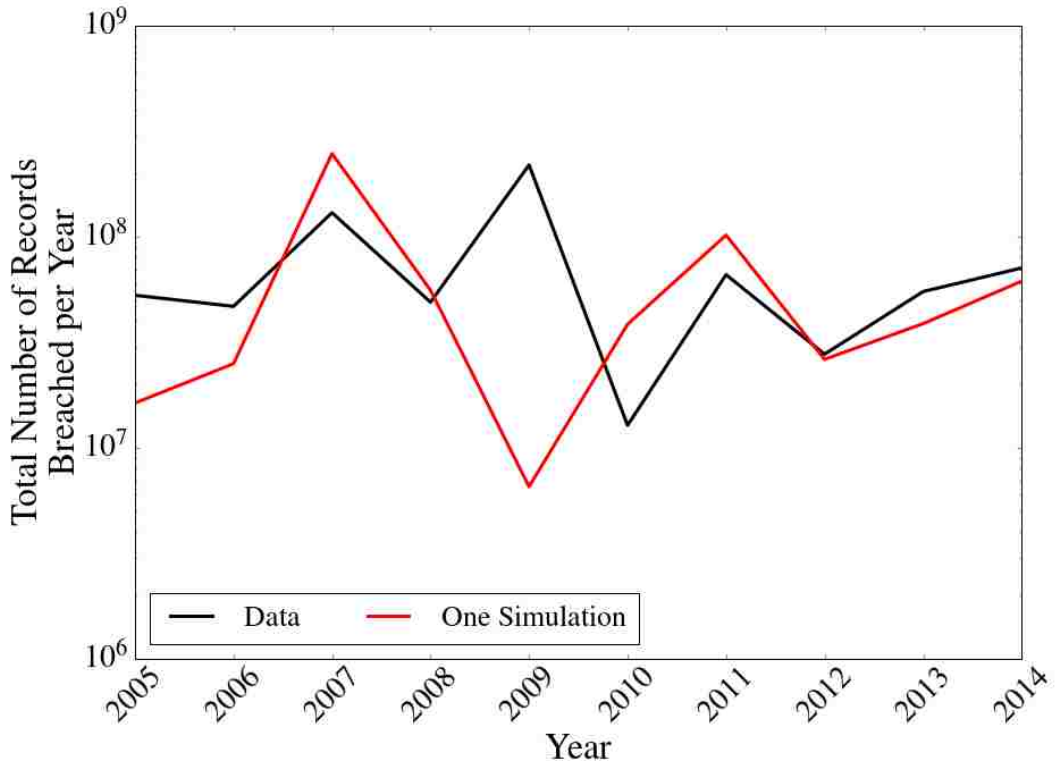

Figure 3.6: The total number of records breached for each year of data along with simulated total sizes of breaches.

We used the models derived from the 2005 to September 2014 data to generate 50,000 simulations of breaches from Sep. 15, 2014 through Sep. 15, 2015. For each day in this simulated timespan we generated a random number of breaches using Equation 3.3, and then for each simulated breach we generated a random breach size using Equation 3.1. We plot the cumulative number of records breached in Figure 3.7.

The mean cumulative number of breached records roughly matches the actual cumulative number of records up to February of 2015, when the Anthem Breach exposed 80 million medical records. In the next six months, Premera/Blue Cross experienced a breach of 11 million health care records, the US office of Personal Management

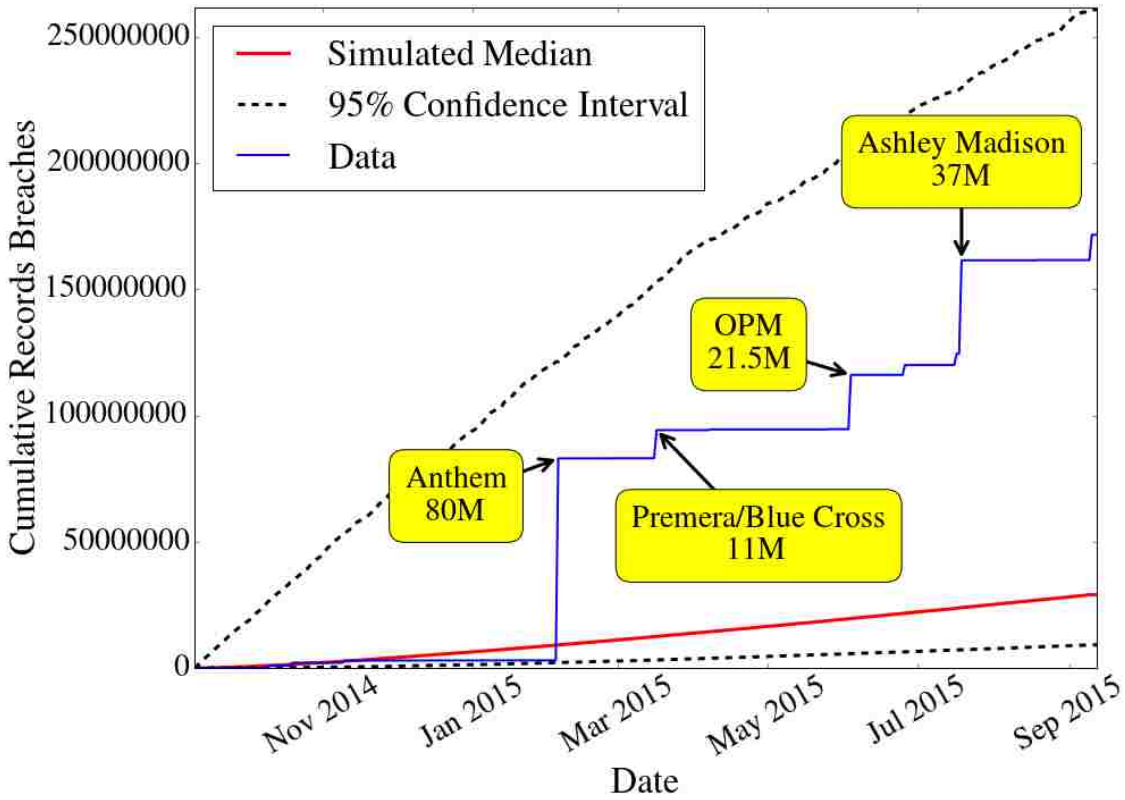

Figure 3.7: The cumulative number of breached records, both historically (shown in blue) and as predicted by our model. The simulated median (shown in red) is computed over 50,000 independent simulations. The dashed lines represent the 95% confidence interval.

experienced a breach containing 21.5 million records, and Ashley Madison experienced the exposure of 37 million user accounts resulting in a significant increase in the total number of records lost. However, all of these breaches are well within the 95% confidence interval of our model.

As discussed in subsection 3.4.1, large data breaches are expected to occur occasionally due to the heavy-tailed nature of the distribution from which they are drawn. However, in our experiments with the model, breaches of the size of the Anthem and Ashely Madison breach occurred in the same year in only 1.08% of

simulations, suggesting that the co-occurrence of these two breach sizes was indeed rare. Although this event was unlikely, it is unclear whether or not it represents a statistically significant change in the overall pattern exhibited by the rest of the data.

3.4.3 Future Breaches

We now use our model built on the past decade of data breaches to simulate what breaches we might expect in the next three years in the United States. With the current climate and concern over data breaches, there will likely be changes in practices and policy that will change data breach trends. However, this gives us an opportunity to examine what might occur if the status quo is maintained. Once again we use the same methodology, predicting from September 15, 2015, through September 15, 2018. We predict the probability of several different sizes of breaches. The results can be seen in Figure 3.8 and Figure 3.9.

Breaches of 1,000,000 records or more are almost certain (99.32%) within the next year. However, in the next year the probability of exceptionally large breaches decreases quickly, with only a 9.77% chance of an Anthem sized breach in the next year. However, in the next three years we can expect to have more large breaches. This is especially clear in Figure 3.9, which shows that we are almost certain to see a breach of 10 million records or more in the next three years (86.2%), but above that size the probability drops off rapidly, e.g. a breach of size greater than 80 million has less than a 25.7% chance of occurring in the next three years.

Predictions like this could be relevant for policy makers interested in the problem of reducing data breaches. For example, the results suggest that it might be more sensible to address the problem of smaller breaches that are almost certain to happen, than to focus on the very large and infrequent headline-grabbing events. Disclosure

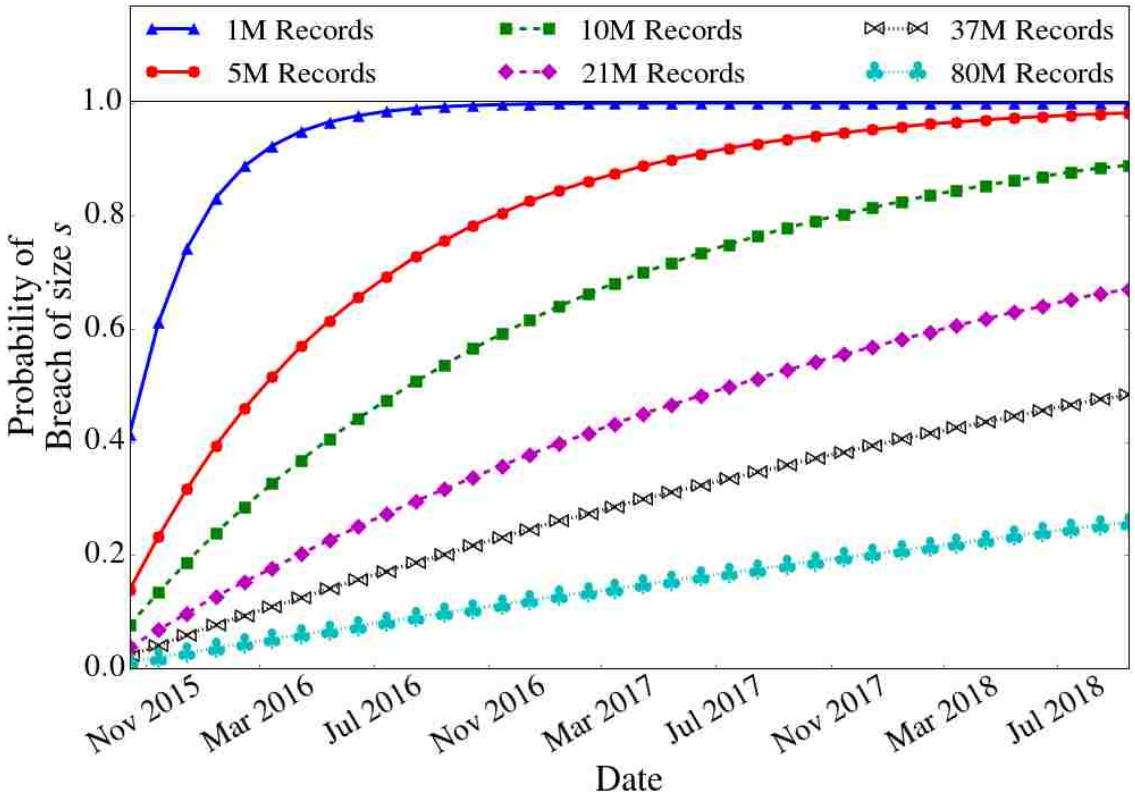

Figure 3.8: The predicted probability of breaches of various sizes over the next three years. Each line represents the probability of at least one breach of the size denoted in the legend occurring before the date on the horizontal axis. We do not include smaller breach sizes, as they will almost certainly occur within the next few months.

laws at the Federal level, that force small, local organizations to consistently report breaches, could be one way of doing this.

As with most efforts to model dynamic, real-world phenomena, we expect the predictions to lose accuracy over time. So although our predictions for the next three years could be off, we expect the model to work better for the short term. As a demonstration, beginning September 15, 2015 we predict the probability of various breach sizes in the next year and the next three years. The exact probabilities are given in Table 3.4. Thus, we can say with high probability (99.3%) that a breach of

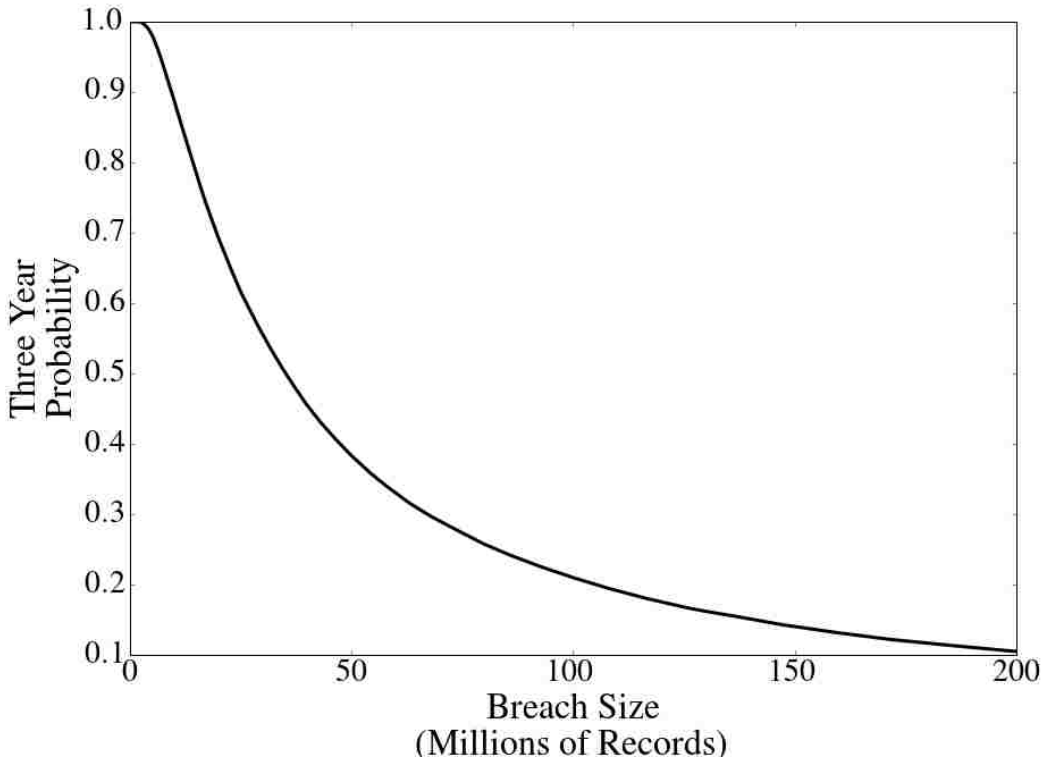

Figure 3.9: The predicted probabilities of breach size after three years.

at least one million records will occur in the next year, and we do not expect to see a breach equivalent to Anthem (9.77% chance). In the next year we expect only a 53.6% chance of a breach of 10 million records or more.

3.4.4 Predicting Future Costs

We can estimate the total expected cost of breaches in the future by incorporating data and other models related to c]st. The Ponemon Institute publishes annual costs of data breaches, and found an average $201 cost per record breached in 2014 [178]. Further analysis by others argues that such a flat rate is not the most accurate

Table 3.4: Chance of the occurrence of various size malicious breaches by in the next year and three years. The breach size is in millions of records.

Breach size (millions)	**% Chance**	
	One Year	**Three Years**
1	99.3	100
5	75.6	98.2
10	53.6	88.9
21.5	31.6	67.0
37	20.1	48.3
80	9.77	25.7
130	5.82	16.2

model for costs. Using non-public data, for example, Jacobs showed that the cost of a breach can be better estimated with a log-log model of the form [130]

$$\log(c) = 7.68 + 0.7584 * \log(s) \tag{3.4}$$

where c is the cost of the breach in data, and s is the size of the breach.

In Equation 3.4 the cost of a breach grows less than linearly, resulting in overall lower costs than those predicted by the Ponemon model. Because the data used to create these models are not public, it is hard to assess their validity, but they illustrate how any cost model can be combined with our results to estimate the future costs of data breaches. Combining these models with Equation 3.1 and Equation 3.3 produces the predicted cumulative cost of data breaches over the next three years, as shown in Figure 3.10.

The flat rate cost model (Ponemon) suggests that in the next three years we can expect anywhere between $8.90 billion and $179 billion in losses associated with public data breaches. Jacob's model gives a more modest estimate of somewhere between $3.87 and $19.9 billion.

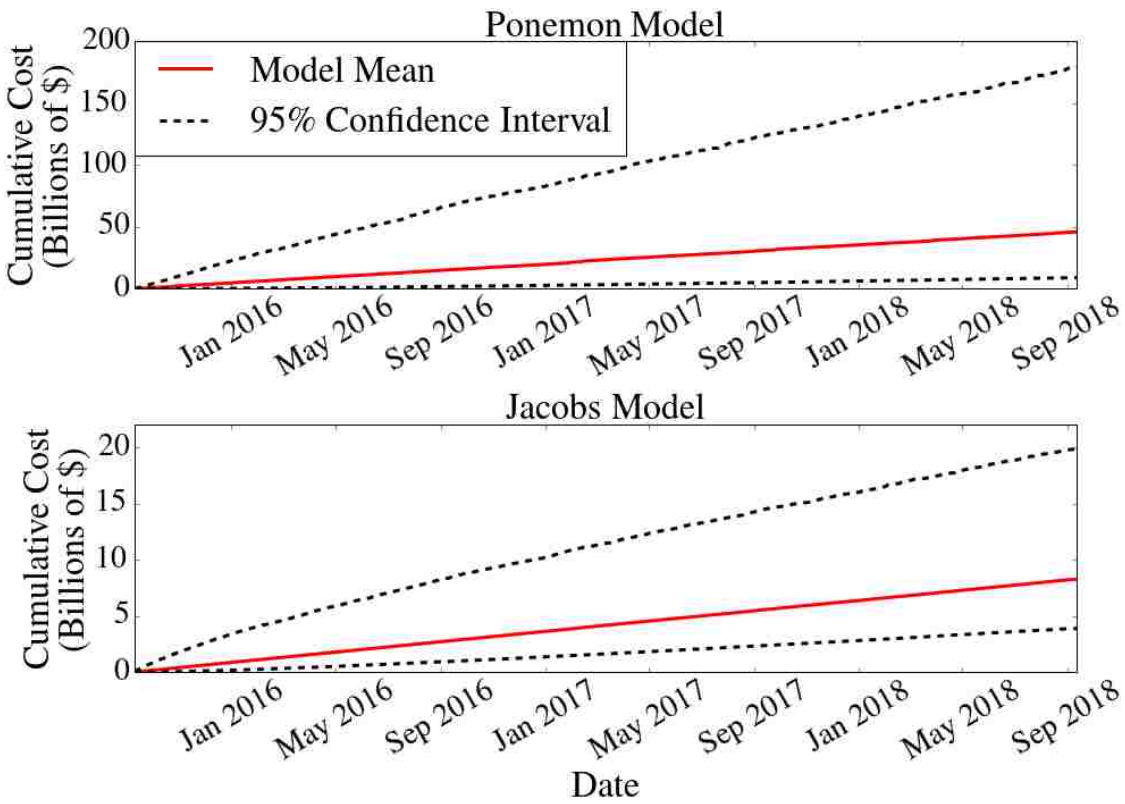

Figure 3.10: Predicted cumulative cost of data breaches in the next three years using two different cost models.

3.5 Related Work

According to the PRC, over 90 reports and articles reference the data used in our study [51]. However, only a few of those reports perform quantitative analysis, and most do not investigate trends in the size or frequency of data breaches. There are a few exceptions, for example, the Symantec Threat Report [55] and the TrendMicro report [124] mentioned earlier. Gemalto reports data breach trends, but do not use the PRC data [96]. Another example is a Verizon report released in 2014 [280], which examines trends in the relative frequency over time of various types of attacks and motivations. However, the methodology for determining the trends is not described,

and the report makes no predictions about the future. Many reports from security companies, such as those from Trustwave [273], focus on classifying the various attack vectors, without attempting to model trends.

Trends in data breaches have received little attention in the academic literature, one exception being Maillart et al.'s analysis of a related dataset [181]. By focusing on the tail of the data their analysis reveals a power-law, which is indistinguishable from the tail of the log-normal distribution we found by considering the entire dataset. Heavy-tailed datasets have also been studied in other domains using similar methods, e.g., [49]. Earlier research investigated trends in the relative frequency of various categories of breaches from 2005-2007 but found that the limited sample size prevented them from making statements about the significance of their results [63]. More recently, in 2010, Widup examined yearly trends in different types of data breaches [290]. However, no statistical analysis was conducted to estimate the underlying distribution or to separate out normal variations from distinct trends. Some papers investigate predictions about future events. For example, Bagchi and Udo developed a general statistical model for predicting the cumulative number security incidents of a specific type [14], and Condon et. al used a time series model to predict security incidents [53]. However neither of these two studies focused specifically on data breaches.

Numerous reports focus on the health care industry. The U.S. Department of Health and Human Services released a 2014 report examining breaches of protected health information [209]. The report includes basic counts of different types of breaches but does not identify any clear trends. Redspin has published three annual reports on data breaches in the healthcare industry [126, 127, 128]. In 2011, they reported a 97% increase in the number of breaches from the previous year, and a dramatic 525% increase in the number of total records breached [126]. The following year, they report an increase in the number of large breaches (22%) and a decrease in

the number of total records breached. These variations fit well with our observations of the heavy-tailed nature of the underlying data.

Some reports focusing on the cost of data breaches were described in subsection 3.4.4. Similar studies focused on hospitals claim that breaches can cost organizations an average of $2.4 million over the course of two years.

Other work has focused on the overall cost of security breaches. Acquisti et al. found a negative impact on the stock value of companies experiencing privacy breaches [2]. Thomas et al. built a branching activity model which measures the impact of information security breaches beyond a breached organization [269]. Studies such as these could be combined with our methodology to infer future overall costs of breaches.

A number of other studies have examined the possible policy implications of data breach notification laws. Picanso suggested a framework for legislation of uniform data breach notifications [218]. Romanosky et al. analyzed the economic and legal ramifications of lawsuits when consumer data is compromised [234]. Later, Romanosky et al. created an abstract economic model to investigate the effect of mandatory data breach disclosure laws [235]. Using older parameter estimates, their model shows that if disclosure were made mandatory, then costs would be higher for companies experiencing breaches and that companies would likely increase their investment in security infrastructure. Graves et al. use PRC data to conclude that credit card companies should wait until fraud occurs before reissuing credit cards in the wake of a breach [108].

3.6 Discussion

Our results suggest that publicly reported data breaches in the U.S. have not increased significantly over the past ten years, either in frequency or in size. Because the distribution of breach sizes is heavy-tailed, large (rare) events occur more frequently than intuition would suggest. This helps to explain why many reports show massive year-to-year increases in both the aggregate number of records exposed and the number of breaches [128, 280, 273, 55, 124, 96]. All of these reports lump data into yearly bins, and this amount of aggregation can often influence the apparent trends (Figure 3.1).

The idea that breaches are not necessarily worsening may seem counter-intuitive. The Red Queen hypothesis in biology [277] provides a possible explanation. It states that organisms not only compete within their own species to gain reproductive advantage, but they must also compete with other species, leading to an evolutionary arms race. In our case, as security practices have improved, attacks have become more sophisticated, possibly resulting in stasis for both attackers or defenders. This hypothesis is consistent with observed patterns in the dataset. Indeed, for breaches over 500,000 records there was no increase in size or frequency of malicious data breaches, suggesting that for large breaches such an arms race could be occurring. Many large breaches have occurred over the past decade, but the largest was disclosed as far back as 2009 [150], and the second largest was even earlier, in 2007 [29]. Future work could analyze these breaches in depth to determine whether more recent breaches have required more sophisticated attacks.

Even if breaches are stable in size and frequency, their impact is likely growing. The ability to monetize personal information, and the increasing ease with which financial transactions are conducted electronically could mean that the cost of data breaches will rise in the future. To address this issue, we considered two different

models taken from the literature, which give wildly different projections. Reconciling these two models is an important area of future work. With improved cost models, however, integration with our models to produce more accurate projections would be straightforward.

Our results are based on publicly available data. It may be that the data are incomplete, and therefore our model is biased downwards, as some breaches will go unreported, but few reported breaches will prove not to have occurred. As more data become available, it will be straightforward to incorporate and update trend analyses and predictions. Given new data, from private sources or other countries other than the United States, it would be important not only to re-analyze trends, but also to revisit the underlying distributions. Despite this caveat, we expect that the PRC data is reasonably complete for the U.S., because most U.S. states already have disclosure laws (48 out of 50 as of January 2015 [213]) that require organizations to report the compromise of sensitive customer information. These laws vary in their requirements so it is possible that many breaches still go unreported. Moreover, different sectors have different reporting laws. For example, the US Department of Health and Human Services requires hospitals to report breaches of medical information containing more than 500 records [1]. This may lead to an over representation of medical breaches in the data. Future work could use interrupted regression to test whether reporting laws change the rate of reporting [284].

As we described earlier, the data are well-modeled by certain distributions, and these distributions could arise from underlying processes related to the breaches (section 3.2). However, Figure 3.2 illustrates that there is some deviation in the tail, suggesting that the log-normal fit is not exact for breaches that exceed 1,000,000 records. There are several possible explanations. It could simply be statistical noise, which is a known consequence of the rarity of large breaches. Alternatively, it could be that large breaches are generated by a different process from smaller breaches,

a hypothesis that we rejected in subsection 3.3.4. Another possibility is that large breaches are more likely to be reported than smaller ones, either because there is a higher likelihood that the breach is noticed or because it is more likely that some of the records are covered by a disclosure law. The negative binomial distribution we observe in breach frequency could be the result of a mixture of different random Poisson processes. Specifically, seem to be reported on different days of the week. Different rates for different organizational types may also explain the shape of the negative binomial distribution.

Different modeling paradigms such those which model large and small breaches differently may result in better predictions. It is also possible that large breaches have become more common very recently, representing a discrete jump in the data, rather than the continuous one used in our models here. Models which account for different days of the week for the frequency of reporting and discrete changes in may provide a better explanation for the data.

This chapter focuses on identifying trends in the size and frequency of data breaches over time, and predicting the likelihood of future breaches. However, it may be possible to identify other factors that influence breaches, for example, the size of an organization. It is reasonable to expect that the number of records that an organization holds is related to its size, and that this factor alone would affect expected breach size. We conducted a preliminary investigation of U.S. universities with breaches in the PRC dataset but found no significant correlation between university enrollments (proxy for size of institution) at the time of the breach and the size of the breach itself. This unanticipated result bears additional study. In the future we plan to identify features of organizations that are predictive of the size and frequency of breaches they will experience, with the goal of helping policy makers focus their attention where it can have the most impact.

Our model provides estimates of the probability of breaches of specific sizes oc-

curring in the past and the future through simulation. Given its relative simplicity, it may be possible to construct analytic solutions for these probabilities, and not have to rely on simulation. However, in general we cannot expect all such models to be tractable analytically.

3.7 Summary

Our analysis of the PRC dataset shows that neither the size nor the frequency of two broad classes of data breaches has increased over the past decade. It is, of course, possible that the PRC dataset is not representative of all breaches or that there has been a significant transition in the underlying probabilities in the recent past which is not yet reflected in our data. A third possible explanation for this surprising result is that data privacy practices have improved at roughly the same rate as attacker prowess—Red Queen effect [277]. Under this scenario, we are in an arms race, and can expect continual pressure to increase defenses just to stay even. It will take extraordinary efforts if we are ever to get ahead.

In conclusion, data breaches pose an ongoing threat to personal and financial security, and they are costly for the organizations that hold large collections of personal data. In addition, because so much of our daily lives is now conducted online, it is becoming easier for criminals to monetize stolen information. This problem is especially acute for individual citizens, who generally have no direct control over the fate of their private information. Finding effective solutions will require understanding the scope of the problem, how it is changing over time, and identifying the underlying processes and incentives.

Chapter 4

Analyzing and Modeling Longitudinal Spam Data[1]

4.1 Introduction

Understanding the longitudinal behavior of security phenomena is only the first step into understanding the impact of interventions. Once we have rigorous analysis of trends, we can begin to study the effect of interventions. For example, botnets have been a persistent security problem but there has been little quantitative analysis of the sustained effect of the most popular intervention to fight botnets. In particular, despite many takedowns in the past decade it is unclear how effective they have been. Simple qualitative observation of declines in malicious activity following a takedown is not sufficient to determine whether the takedown is effective or causal [102, 139,

[1]The substance and much of the material in this chapter was previously published as "Analyzing and Modeling Longitudinal Security Data: Promise and Pitfalls," which appeared in the *Proceedings of the 2015 Annual Computer Security Applications Conference*. Sections 4.4.4 and 4.4.5 have not been previously published. Michel van Eeten provided the spam data, I conceived, designed, and executed the research. My co-authors advised and aided in the preparation of the final written presentation.

144]. Attributing cause is always problematic, but it is especially difficult when empirical datasets have high variance as is often the case in security[78, 74].

In this chapter, we explore some of the opportunities and impediments to analyzing longitudinal security data, by focusing on the concrete example of spam, developing statistical models to describe a large dataset, and using the model to assess the effect of certain interventions. We ask whether a particular intervention has a temporary or sustained impact and how interventions play out geographically. A potential pitfall in longitudinal datasets, including our dataset, is high variance, and we use careful statistical methods to separate significant effects from noise. A second issue is the retrospective nature of data-driven analyses, which makes predicting the future a challenge. Because intervention methods are often re-used, however, we believe that studying the existing examples, e.g., a historical botnet takedown, can provide insight about the likely effect of similar future interventions.

We illustrate our approach by analyzing a spam dataset, comprising more than 127 billion spam messages sent from over 440 million unique IP addresses, spread across 260 ISPs in 60 countries. Spam is a global problem, and countermeasures have never eliminated it completely. Spam plays a key role in the cyber-crime ecosystem as a vector for various activities such as stealing login credentials through phishing, distributing malware, making fraudulent sales, or selling illegal goods [151]. Spam can be viewed as a proxy for estimating the numbers of infected PCs and the extent of botnets [312, 267, 132].

To compare spam levels across countries, we study a quantity called *wickedness* [118], which can be thought of as the concentration of infected machines sending out spam, either in a single Internet Service Provider (ISP) or in a geographic region. This measure allows us to compare spam levels among different countries or different ISPs, identify how different factors contribute to the concentration of spam sending computers, and assess what effect interventions have across the globe.

Analysis of the data shows that spam concentrations are relatively stable for ISPs from one week to the next but are punctuated by spikes that often span several orders of magnitude. These spikes can mask the effect of interventions. Further analysis also reveals that: (1) Gross Domestic Product (GDP) per capita is negatively correlated with wickedness, with less developed countries experiencing higher levels; (2) an ISP's wickedness is correlated with that of surrounding ISPs, suggesting that there are regional influences; and (3) an ISP's network connectivity is correlated with wickedness.

To further understand the impact of ISP connectivity on spam, we construct an *ISP graph* that represents how ISPs are connected to each other. The graph reveals that ISPs with high graph centrality have lower wickedness, while those on the periphery suffer higher rates of infection. Adding a simple model of spam dynamics to the ISP graph shows that spam concentrations at an ISP are influenced by previous levels, suggesting that spam could is one driver in spreading infections across the Internet.

Using these observations we constructed a number of models. We tested many and adopted the simplest statistical model that performed well on our dataset: autoregression, which uses past wickedness and the external factors mentioned above to predict current wickedness. Despite its simplicity, this approach outperformed several alternatives such as: Support Vector Machines, Decision Trees, Artificial Neural Networks and Gradient Tree Boosting. In addition to its autoregressive component, our model incorporates five other relevant factors which we found improved the model's explanatory power: national economic indicators, regional wickedness levels, Autonomous System (AS) topology, and traffic flows on the Internet. An important aspect of the model is the structure of the ISP-level network, which influences wickedness. For example, an ISP on the periphery generally has higher observed levels of wickedness than one in the core. We measured the impact of these effects

and incorporated them into the model.

In the last decade, a number of approaches have been suggested and implemented to help fight spam. Of these, the most famous is the botnet takedown. But, email providers have also adopted adaptive IP black lists [105], banks have restricted access to credit card payment processors [136], resources have been devoted to arresting and prosecuting cyber-criminals [192, 149, 5], and users of infected computers have been offered free cleanup tools [30]. Some of these interventions seem to have led to declining spam levels, e.g., real-time filtering and credit card interventions [268, 189, 136, 265].

We show how modeling can help identify when particular interventions likely began affecting spam concentrations. The best model of our dataset identifies three distinct time periods or *eras*, each corresponding to different dynamics. These eras correlate roughly with the introduction of new intervention strategies, and they give some idea of the overall impact of a particular strategy.

When the exact date of an intervention is known (as in the case of botnet takedowns), we can use the model to analyze its impact more precisely, both globally and regionally. Model analysis confirms the hypothesis that most botnet takedowns are effective only in the short-term, with spam levels rebounding in the weeks after a takedown [153]. However, we also find that a few of the takedowns were globally effective in the long term. A closer look at their regional impact, however, shows that effects vary dramatically across different geographic areas and individual countries. Takedowns that are successful globally can be detrimental in specific countries.

Our work uses one particular dataset to illustrate how robust statistical techniques can be applied to study spam trends and the effect of interventions—globally, regionally, and by individual country. Because we studied data taken from a single data source, and focused only on email spam, our conclusions are only as good as the

data—a pitfall of any statistical analysis. The methods, however, could readily be applied to other sources of spam and even other security data, as they become available. Additional datasets would certainly improve our confidence in the conclusions of the analysis, and section 4.2 discusses the idiosyncrasies of our particular dataset.

In summary, statistical analysis of global longitudinal data is a promising approach to understanding the security landscape. This chapter makes the following contributions:

1. It presents a robust statistical analysis of longitudinal, global security data, showing how to analyze high variance time series, identify correlations with external factors, and identify the effects of interventions, both when the deployment date is unknown (filters) and when it is known exactly (botnets).

2. It identifies statistically significant correlations between spam concentrations and various risk factors, including GDP, nearby spam concentrations, and ISP connectivity in the ISP graph. Traffic dynamics on this graph influence future wickedness, suggesting that spam is used to spread malware infections.

3. Identification of three statistically distinct eras within the ten-year data set. Although spam levels are highly variable in all eras, the overall concentration of spam declines during the last two eras. These declines may be related to historical events that are outside the scope of our study, and they may have caused discernible shifts in the data.

4. Analysis of the global impact of historical botnet takedowns: only a few of the studied takedowns had lasting impact, while most had only a transient effect, in all eras.

5. Geographic impacts of takedowns. We find that even when a takedown is effective globally, it often results in an increase in wickedness in particular

regions or countries.

4.2 Collecting and Mapping Spam Data to Wickedness

In this section we describe our dataset, and the *wickedness* metric. We show that wickedness has interesting statistical properties, and identify significant changes in wickedness over time.

4.2.1 Spam Data

Our spam dataset is based on that used by Van Eeten et al. [276] but greatly expanded. We collected additional data, doubling the timespan covered, and studied the data on a weekly basis. The original study examined spam trends only on a quarterly basis. This dataset was collected from a *spam trap*—an Internet domain designed specifically to capture spam with e-mail addresses that have never been published or used to send or receive legitimate email. Spam traps have been used successfully to identify malware infected hosts, and to measure the extent of botnets, because botnets often send spam [312, 267, 132]. Over the past decade, our spam trap received more than 127 billion spam messages, sent from 440 million unique IP addresses.

In order to make comparisons among different ISPs and geographic regions, the ISP which owns each IP address and the country in which that ISP operates must be identified. To do this we used the following procedure:

1. Each IP address was linked to an ASN (Autonomous System number) using historical BGP data.

2. Each ASN was then manually linked to an administrating entity using historical WHOIS records.

3. Industry reports and news media were consulted to connect the administrating entities to the main ISPs in 60 countries, as identified in Telegeography's GlobalComms database. The database also provides us with accurate subscriber numbers for each ISP.

4. Each (part of an) ASN was mapped onto a country using MaxMind's GeoIP database [177].

The manual mapping of ASNs to ISPs prevented us from identifying all possible ISPs which sent spam to our trap. However, we were able to map 659 ASNs to 260 ISPs in 60 countries. These ISPs account for over 80% of the major broadband markets in those countries. These countries also compose the entirety of the Organisation for Economic Co-operate and Development(OECD) and European Union, along with several other major spam sending nations.

This procedure produced two time series for each ISP: a count of spam messages and the number of unique IP addresses that sent spam per day. Some ISPs provide dynamic IP addresses with short lease times to their customers. This could lead to a single infected host being associated with two IP addresses. To help correct for this potential source of overcounting, we use average daily counts of IP addresses over the course of a week to obtain an estimate of the number of infected machines associated with an ISP in a given week. This produces slightly coarser granularity data but removes some of the churn caused by dynamic addresses.

Our data was collected from a single spam trap and is only a sample of all the spam sent globally, and it is possible that our data reflects the activity of only a few unsophisticated spam gangs. It is difficult to exactly compare our data to other publicly available spam reports because most reports rely on relative measures

such as fraction of total email that was classified as spam or percentages relative to a peak. However we were able to make some qualitative comparisons to other sources. A subset of the data from 2006 and 2009 was previously found to correspond with industry reports, both in terms of spam volume over time and geographical distribution of sources [276].

Comparing post 2010 trends to longitudinal data available from Spamhaus [256], our data on global wickedness qualitatively matches theirs until mid-2012. After that, however, Spamhaus shows a brief rise in spam, though not to previous levels, while our data show a continued downward trend (Figure 4.2). Symantec reports a small overall decline in annual average spam in 2013 [266], and Kaspersky also reported a small decline in the percentage of spam email compared to legitimate email in 2013 [110]. Our data also shows declines in these two years. The discrepancies between our data and Spamhaus likely reflect changes in tactics of spammers over time that are not captured by our spam trap. However, in this chapter we emphasize the procedure used to analyze the data over the exact conclusions drawn from the analysis, which in future work could be verified by analyzing other datasets.

4.2.2 Estimating Wickedness From Spam Data

We calculate wickedness in terms of IP addresses sending spam. The two time series establish the total number of spam sending hosts within an ISP, but they do not account for the total number of IP addresses actively used by each ISP, i.e. the number of customers. We focus on wickedness rather than the absolute number of spam sending hosts to allow valid comparisons between ISPs, countries, and regions in the world. We use data from TeleGeography's Globalcomm database to establish the number of subscribers for each ISP. These data, available quarterly, allowed us to compute the concentration of malicious hosts per customer (the wickedness) and

the number of spam messages sent per customer.[2] Using linear interpolation, we inferred the number of customers each week to match the time granularity of the data for malicious hosts. We calculate the wickedness of an ISP i at time t as:

$$W_i(t) = \frac{A_i(t)}{C_i(t)}. \tag{4.1}$$

where $A_i(t)$ refers to the number of spam-sending IP addresses and $C_i(t)$ refers to the number of customers for ISP i at time t. The *global* wickedness is defined over all ISPs, i.e. $W(t) = \sum_i A_i(t) / \sum_i C_i(t)$. Figure 4.1 and Figure 4.2 shows the global wickedness over time calculated from our dataset.

In these data, which capture a sample of the total population of spam-emitting hosts worldwide, between 0.00091% and 0.33% of hosts are sending spam at any given time. However, individual ISP infection rates vary widely as shown by the shaded area in Figure 4.1, with some ISPs as high as 80% and others with 0%. Moreover, a single ISP's infection rate often varies by several orders of magnitude from one week to the next. For example, in April, 2011 an ISP in Pakistan experienced a more than 800-fold increase in wickedness in a single week. Previous work has also observed highly dynamic infection levels in IP space [41, 237].

In spite of this large variation, our analysis shows that wickedness at the individual ISP level is highly *autocorrelated*, i.e. the correlation between wickedness in any given week and the previous week is high (Kendall's $\tau = 0.93$). Kendall's τ is a non-parametric measure of statistical dependence. Unlike the more widely used Pearson's r, Kendall's τ does not assume a linear relationship between the data, and is therefore better able to identify non-linear relationships, which abound in our data [140].[3] This counterintuitive result is explained by the fact that in the vast

[2] Alternatively, wickedness could be defined using messages per customer. We have analyzed the data both ways (data not shown), with essentially identical results.

[3] Measures of linear correlation between the $\ln W_i(t)$ and $\ln W_i(t-1)$ are exceptionally

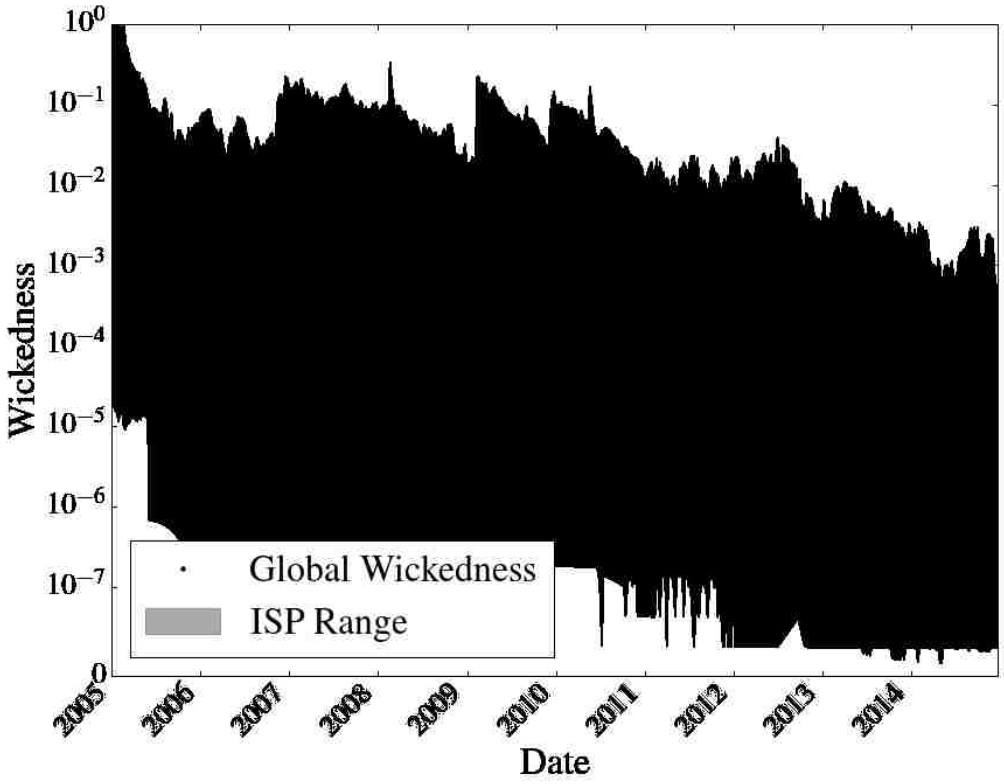

Figure 4.1: Global spam viewed on a log vertical axis to show the high variance in the spam data. Black points indicate global wickedness, and the shaded area shows the range of values for individual ISPs.

majority of cases week-to-week variation is small, even though a minority of cases break this pattern by varying over several orders of magnitude. Such high variance can often lead to erroneous conclusions about data. Many statistical methods require that data have limited variance, and using such methods might indicate significant changes when none exist [74].

Figure 4.2 shows several possible qualitative changes in spam volume, and in subsection 4.4.2 we find that spam exhibits statistically significantly different behavior

high (Pearson's $r = 0.990$), suggesting a nonlinear relationship similar to a power law. This informs the construction of our model in section 4.4.

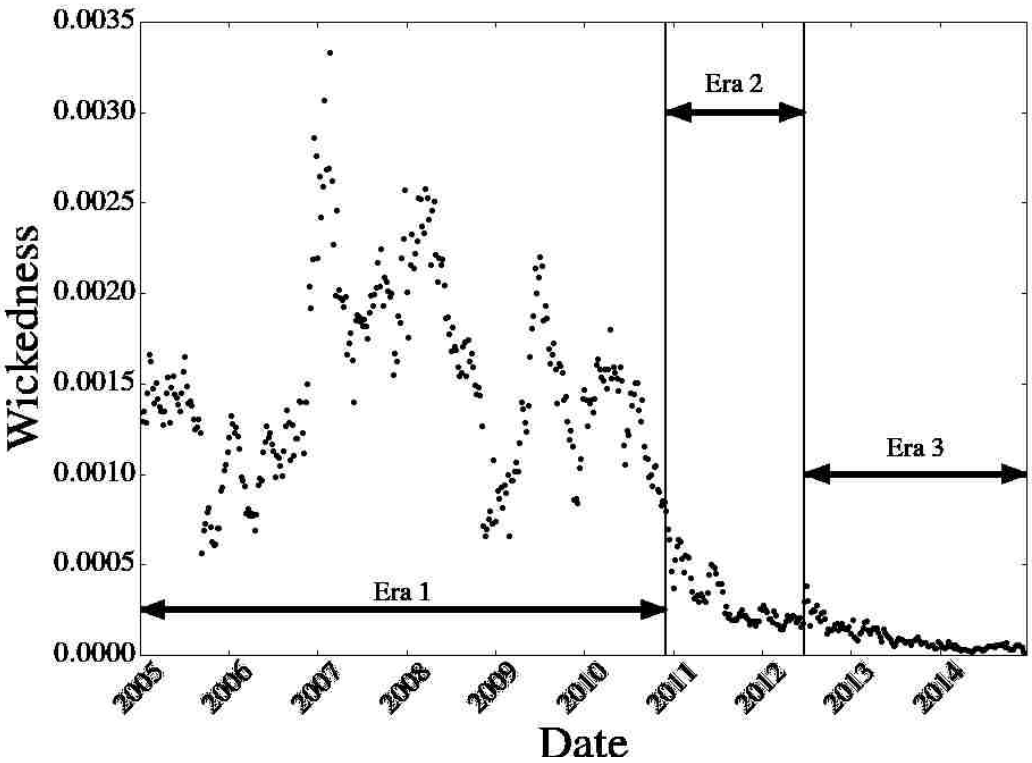

Figure 4.2: Global spam viewed with a linear vertical axis to show the qualitative changes in wickedness between different eras.

during these periods.

Era 1: Beginning in 2005, spam increased dramatically until the botnet takedowns began in 2008. During this era spam levels were volatile, punctuated by sharp increases and decreases both globally and at the ISP level.

Era 2: In mid 2010, spam levels began to drop dramatically. We find a statistically significant effect in late 2010.

Era 3: In mid 2012, a spike is observed in the data, followed by further decline in wickedness. The variance in global wickedness also decreases.

These three eras are highlighted in Figure 4.2. In subsection 4.4.2 we use maximum likelihood techniques to pinpoint when statistically significant transitions occurred and discuss possible causes of these transitions.

4.3 Risk Factors

The previous section defined wickedness and examined its properties in our dataset. Next we ask if certain external "risk factors" are related to an ISP's level of wickedness. In this section, we consider demographic factors, the effect of geography, network effects, and traffic dynamics.

4.3.1 Demographic Factors

Previous work identified correlations between spam concentrations and measures of development, such as Internet use per capita or education [276, 308]. We find similar results using gross domestic product per capita (GDP). GDP data were obtained from the World Bank, which produces annual data on a per-country level for multiple demographic factors [16]. We use GDP per Capita because recent data is readily available, but other measures of development such as unemployment or corruption within institutions might also be instructive. We used linear interpolation to infer weekly values from the annual data.

For each week of data, we compute τ between ISP wickedness and the GDP of the country in which each ISP was operating. The top panel of Figure 4.3 shows these correlations over the course of 520 weeks, and indicates that GDP is consistently negatively correlated with wickedness, in agreement with results from previous studies [276, 308]. In subsection 4.4.1 we calculate the size of this effect. The correlation decreases in the later portions of the data, which could indicate that infection rates

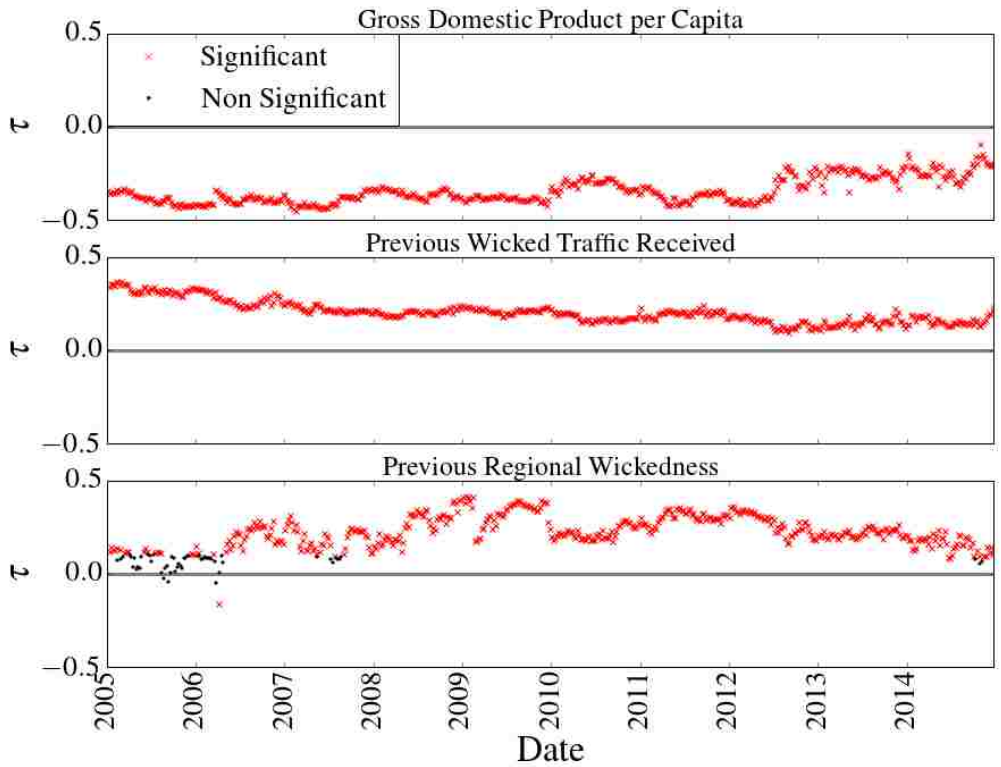

Figure 4.3: Correlation between wickedness and GDP (top panel), wickedness and traffic (middle panel), and wickedness and average regional wickedness (bottom panel). The vertical axis in all plots is Kendall's τ between wickedness and traffic during the week shown on the horizontal axis. Red indicates significant correlations at the $p < .05$ level.

are becoming less tethered to development, as technology levels rise across the globe.

4.3.2 Geographic Clustering

Qualitatively, we observe that wickedness levels cluster in certain geographic regions during specific periods. For example, during January, 2011 high levels of wickedness are observed in Eastern European countries. Roughly a year later, wickedness

declined in Eastern Europe but increased in Southeast Asia[4].

To study this geographic clustering, we divide the world into 14 regions, defined by the United Nations [204], and measured the correlation between the wickedness of an ISP and the average wickedness of all other ISPs in the same region (excluding the original ISP) in the previous week.

We find significant positive correlations between this value and wickedness throughout most of the data (see Figure 4.3). We study this result more in depth in section 4.4.

4.3.3 Autonomous System Topology

Another possible risk factor is an ISP's position in the topological structure of the Internet at the Autonomous System (AS) routing level. To investigate the strength of this effect, we measured the correlation between wickedness and several popular topological metrics [77]. This is not straightforward because our data were collected at the ISP level, and connectivity between ISPs is not identical to Autonomous System connectivity. We address this problem by constructing a hybrid network that reflects both topologies.

We constructed this new network by beginning with the AS level, retrieving AS network data from the *Internet Research Lab's Internet AS-Level Topology Archive*.[5] The archive collects daily and monthly snapshots of AS-level topology from a number of different sources and, at the time of download on February 11, 2015, was one of the most complete publicly available sources of the AS-level Internet topology [214]. We construct the ISP graph using the following steps:

[4]Map not shown due to space constraints.

[5]`http://irl.cs.ucla.edu/topology/ipv4/daily/`.

1. *Aggregate nodes*: Combine all ASNs owned by a single ISP into a single node. This produces a graph that contains both ISP and ASN nodes.

2. *Aggregate edges*: If there are multiple edges between two nodes, combine them into a single weighted edge, with weight equal to the number of connections between the nodes.

3. *Remove stubs*: Remove ASN nodes that are not directly connected to an ISP and have degree equal to one.

4. Combine the daily version of the graph into a weekly snapshot by taking the graph union.

We remove stub ASes because they likely have little real-world influence on traffic flow in the ISP graph [180].

Using this hybrid graph, we investigated the correlation between ISP wickedness and a number of popular measures of graph topology [77]. In total we tested eight different measures.

Figure 4.4 shows the correlation between wickedness and the six of the eight topological features we tested. Three features are significantly correlated with wickedness throughout the study period (top two panels, and middle right panel): an ISP's location within the Internet hierarchy (Core Number and Average Shortest Path Length) and centrality (weighted degree). Weighted degree is correlated for the majority of time steps (middle right panel), excluding the early part of the time series, a few weeks in 2010 and 2011, and late in the data. By contrast, betweenness centrality and clustering coefficient do not show significant correlation throughout the time series, while page rank is correlated roughly one third of the time. The correlations that do exist show that in general ISPs with high centrality (degree), tend to have low wickedness, while ISPs on the periphery of the network (low core number, high

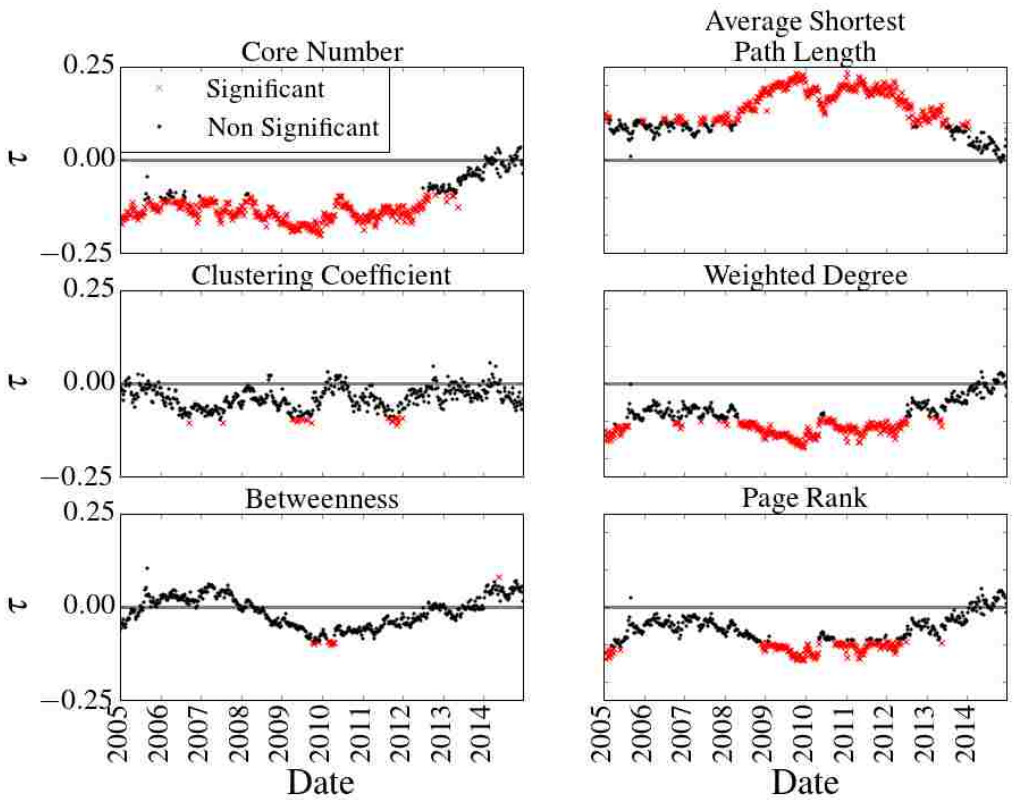

Figure 4.4: Correlation between wickedness and ISP graph topology. The vertical axis in all plots shows the Kendall's τ between wickedness and the topological measure for the corresponding week on the horizontal axis. Red indicates significant correlations at the $p < .05$ level.

average shortest path length), have higher wickedness values. It is not clear why this is the case; one possibility is that ISPs on the edge of the network tend to be smaller and thus have fewer resources to counter infections.

ASNs are often categorized by the type of services they provide [69], and this could influence their level of wickedness. We did not include this factor in our analysis because each ISP could be an aggregation of multiple ASNs, making clear categorizations of the services provided by an ISP difficult to ascertain. Moreover, since our data on subscriber numbers is at the ISP level, we cannot easily allocate it

to different ASNs.

4.3.4 Network Traffic Dynamics

Traffic dynamics affect the concentration of malicious hosts [118], but appropriate network traffic datasets are not publicly available. Numerous models of traffic flow have been proposed for the AS network, ranging from simple [236] to elaborate [37], and for this study we adapted Roughan et al.'s gravity model [236] to simulate malicious traffic between nodes in the ISP graph. In the gravity model, the traffic received by node i from j is expressed as:

$$r_{ij} = \frac{C_i C_j}{d_{ij}^2} \tag{4.2}$$

where C_i is the number of customers for ISP i, and d_{ij} is the shortest path length between the two ISPs in the ISP graph. We assume that malicious traffic is proportional to the total traffic received by an ISP i, and then calculate the expected per customer rate of malicious traffic:

$$R_i = \frac{\sum_{j \neq i} R_{ij} W_j}{C_i} \tag{4.3}$$

where W_j is the concentration of spam-emitting IP addresses at ISP j and R_{ij} is fraction of j's traffic destined for i (normalized r_{ij}). Normalizing by C_i allows us to interpret R_i as the expected fraction of malicious traffic received by each customer of ISP i.

We test whether this calculated value correlates with wickedness in the same way we did for the topological factors, except we consider time by introducing a one time-step lag between the two series. This allows us to identify possible causal

relationships between traffic and wickedness[107], as shown in the top panel of Figure 4.3. The figure shows that there is a statistically significant positive correlation through time. This indicates that the flow of malicious traffic, in particular, the amount of malicious traffic received per customer in the previous week, correlates with increased wickedness in the next week.

4.4 Modeling

In the previous section we identified external factors that are individually correlated with wickedness. In this section, we develop an autoregressive model that incorporates and combines these factors. We then use the model to explore the relative strengths of these effects and identify the transitions between spam eras.

We also evaluate the accuracy of our model compared to several more complex and sophisticated alternatives. Using the results from section 4.2 and section 4.3, we tested numerous models [26], including: random walk models, more complex autoregressive models, support vector machines, feed-forward neural networks, decision tree regression, and gradient tree boosting. Surprisingly, the simple autoregressive model presented in subsection 4.4.1, fits the data as well as more complex models and has greater explanatory power.

4.4.1 Autoregressive Model

An *autoregressive* model is a type of linear regression, which uses previous values in a time series to predict future values. We have already discovered in subsection 4.2.2 that our dataset is highly autocorrelated, which justifies this model selection, and we include the external risk factors identified in section 4.2.

Visual inspection of the data reveals an obvious decline in wickedness levels somewhere after 2010. We incorporated this observation into the model by hypothesizing up to three distinct temporal eras. In each era y, the wickedness of ISP i at time t is modeled as:

$$\begin{aligned}\ln(W_{i,y}(t)) =& \beta_{0,y}\ln(W_i(t-1)) + \beta_{1,y}\ln(R_i(t-1)) + \\ & \beta_{2,y}\ln(G_i(t-1)) + \beta_{3,y}\ln(E_i(t)) + \\ & \beta_{4,y}P_i(t) + \beta_{5,y}\ln(D_i(t)) + \epsilon_y\end{aligned} \tag{4.4}$$

Each symbol in Equation 4.4 is described in Table 4.1. In subsection 4.3.3 we found that both average shortest path length and core number are correlated with wickedness. However, these two measures are highly correlated with each other, and including both metrics in the model could cause estimates of $\beta_{x,y}$ to be incorrect [298], so we selected average path length.

All autoregressive models include a distribution of error terms, here represented by ϵ, and they are usually assumed to be normal [298]. In our case, given the high variance of the data (section 4.2), we assume $\epsilon_y \sim T(\nu, \sigma)$, where $T(\nu, \sigma)$ is the non-standardized Student's T distribution, which is considered to be more appropriate when a dependent variable has high variance [298] which we observed in section 4.2.

In the model some variables are log transformed because preliminary inspection revealed that their functional relationships were non-linear in particular ways (i.e. roughly linear on log/log plots).[6]

[6]We speculate that the log/log relationship between $W_i(t)$ and $W_i(t-1)$, may arise from an underlying growth or decay process in malware infected hosts.

4.4.2 Identifying Model Transitions

In section 4.1 we noted that the data appear to experience qualitative changes, which might correspond to changes in spam tactics or the development of new spam fighting tools. However, it is unclear exactly when these changes might have occurred. Rather than pre-define transitions between eras based on industry reports or qualitative evaluations of the data, we used the model to determine the most likely dates when significant changes in spam concentrations occurred, testing for zero, one, or two significant transitions.

For each possible combination of two transition dates, we use maximum likelihood estimation (MLE) to estimate the values for all $\beta_{x,y}$ and their standard errors. We then selected transition dates which gave the model the highest likelihood.

To measure whether dividing the data into three eras is justified, we compared the model to one with a single division into two eras, and one with no divisions. We used the Akaike Information Criteria (AIC)[298], which is a measure of goodness of fit based on likelihood that penalizes more complex models. We found a statistically significant improvement between the model with two divisions ($AIC = 14403.3$) and models with a single($AIC = 15709.0$) or no divisions($AIC = 27276.8$)). It is also possible that there are more statistically significant transitions in the data than we were able to test for due to computational constraints. We leave this topic for future investigation.

The first change identified by our methodology begins in December 2010, after which we see a steady decline in spam levels. This may be due to improvements in adaptive, real-time filtering, which were first deployed at companies such as Google as early as 2006 [268]. There is evidence that improved filtering forced spammers to adopt new more costly methods of spamming, such as large-scale account hacking [105]. Filtering even affected delivery of legitimate bulk email in the first half

of 2011 [217]. Microsoft's Security Intelligence Report attributes the decline in 2011 to both more sophisticated filtering techniques, and to the takedown of the Cutwail and Rustock botnets [189].

We identify a second transition beginning in June 2012. In May, 2011 Kanich et al. published a paper which identified a handful of banks that were responsible for processing most of the payments made by spam victims [136]. Shortly after the paper was published, Visa tightened requirements for merchants, and effectively disrupted many spammers' revenue streams [282]. Seven months after the announcement of these requirements, spammers reported difficulty maintaining reliable credit card processing [155] and spam volume dropped significantly, e.g. Symantec's Internet Security Threat Report from 2012 notes a significant drop in pharmaceutical spam [265].

4.4.3 Model Results

Table 4.1 gives the MLE values for the $\beta_{x,y}$. Examining Table 4.1 we see that the autoregressive term has the largest influence on future wickedness. Surprisingly, one of the other terms (regional wickedness during the previous week) in all eras has an opposite effect from what was reported in section 4.3 (Figure 4.3). This is an example of *Simpson's Paradox* [232], indicating that in the presence of other variables, high levels of wickedness in neighboring ISPs actually reduce future wickedness. One possible explanation is that spammers initially try to infect as many machines in a region as possible, and then concentrate on vulnerable ISPs as they discover them, reducing attacks on the less vulnerable ISPs. This factor and the other variables identified in section 4.2 are statistically significant, but at low levels. This simple model accounts for the vast majority of the variance in our data, with a combined

coefficient of determination of $R^2 = 0.980$ for data in all eras.[7] It is possible that more sophisticated models might provide more predictive power than our simple linear, autoregressive model. We tested support vector machines, feed forward neural networks, decision tree regression, and gradient tree boosting, and found that none outperformed our model (measured by R^2) or had similar explanatory power. Moreover, our robust statistical approach can determine statistical significance without computationally expensive procedures, such as cross validation.

4.4.4 Cross-validation

At this point we have a best fit model to the existing data. One way to assess the validity of the model is through repeated random sub-sampling cross validation [143]. To do this, we partitioned the dataset by randomly assigning $\frac{2}{3}$ of the ISPs to the training set and the remainder to the testing set. Then we re-estimated the $\beta_{x,y}$ values using only the training data and used this re-calculated model to predict wickedness for the training and testing data separately, measuring accuracy with R^2. We iterated this process for 10,000 repetitions.

There was no statistically significant difference in mean R^2 values between the training and testing data (Wilcoxon signed-rank test[292] $p > 0.1$), and they differed from the R^2 on the complete data set by less than 0.001%. The mean $\beta_{x,y}$ values from the cross-validation were not statistically significantly different than the $\beta_{x,y}$ values calculated in Table 4.1 (one sample Student's T-test[262] $p > 0.1$). We also cross-validated the model by subsampling in the time domain (selecting $\frac{2}{3}$ of the weeks for training and assigning the rest to testing), with similar results. These results are not surprising because generally the standard errors of regression coefficients correspond

[7] The autoregressive term is mostly responsible for the high R^2 in the model. However, without the autoregressive term the model still has an $R^2 = 0.58$, indicating moderate explanatory power.

Table 4.1: Coefficients for the autoregressive model. Range indicates the range of possible values for each variable. **Bold** coefficients are statistically significant at the $p < 0.01$ level.

Variable	**Symbol**	$\boldsymbol{\beta_{i,y}}$	**Era 1** Jan 2005-Dec 2010	**Era 2** Dec 2010-June 2012	**Era 3** June 2012-Dec 2014
Log Prev Wickedness	$\ln(W_i(t\text{-}1))$	$\beta_{0,y}$	**0.994**	**0.991**	**0.976**
Log Prev Wicked Traffic	$\ln(R_i(t\text{-}1))$	$\beta_{1,y}$	0.0002	0.0003	**0.0145**
Log Prev Region Wickedness	$\ln(G_i(t\text{-}1))$	$\beta_{2,y}$	**-0.0039**	**-0.0158**	**-0.0188**
Log GDP per capita	$\ln(E_i(t))$	$\beta_{3,y}$	**-0.0080**	**-0.0255**	**-0.0359**
Shortest Path Length	$P_i(t)$	$\beta_{4,y}$	**0.0052**	0.0109	**0.0658**
Log Weighted Degree	$\ln(D_i(t))$	$\beta_{5,y}$	-0.00009	-0.0006	**0.0175**
		$\boldsymbol{R^2}$	0.985	0.975	0.937

to the distribution of values from such a cross-validation. However, these results emphasize that our results are robust under multiple tests of validity.

4.4.5 Alternative Models

Although the simple auto-regressive model is surprisingly effective, certain properties of the model suggest that alternatives might be even better. Autocorrelation was observed in the residuals of the model indicating that using additional autoregressive terms could improve the fit. The Akaike Information Criterion (AIC), used to measure the balance between accuracy and parsimony of statistical models, can be used to select among possible autoregressive models [32]. Using AIC we found that the single one-week lag term provides the best possible autoregressive model.

The fact that the single autoregressive coefficient is near 1.0 suggests that the data might be well-modeled by an even simpler model, a random walk. We tested this hypothesis with several different kinds of random walks, each of which produced unrealistic results (not shown) and failed tests designed to identify random walks [70]. Other diagnostic tests were used to evaluate the autoregressive model and showed that the model coefficients are robust to common pathologies of time-series models, such as heteroskedacity and residual autocorrelation [32], and that the residuals fit the assumed student's T distribution.

More complex models [26] such as support vector machines, feed forward neural networks, decision tree regression, and gradient tree boosting might be expected to outperform our linear, autoregressive model. These models generally achieve higher predictive accuracy, at the cost of less explanatory power. Thus, it can be challenging to find definable relationships between endogenous and exogenous variables in these more complex models. We fit the data to all four of the aforementioned models. Surprisingly, even on predictive accuracy, where these models excel, none

had statistically significantly better performance, as measured by R^2. This suggests that the structure of Equation 4.4 is likely correct and that the residuals arise from unknown external factors that can be treated as statistical noise.

4.5 The Effect of Takedowns

Section 4.4 presented a statistical model that accurately assesses the relative contribution of a variety of factors on spam levels over almost a decade. This section shows how the model can be used to study the impact of interventions such as botnet takedowns.

Although spam levels typically drop immediately following a takedown, there is anecdotal evidence that this effect is short-term, often returning to previous levels within a few weeks [168, 281, 229]. Given the high variance in the data, however, quantifying the short-term and long-term effects is challenging, and requires rigorous statistical testing. With only a small extension to the model, we can conduct such tests and consider the impact of takedowns on different regions of the world.

4.5.1 Modeling Takedowns

We model takedowns, which are a discrete event at the timescale of our data, by adding binary variables to the model:

$$B_k(t-j) = \begin{cases} 1 & \text{takedown } k \text{ occurred } j \text{ weeks ago} \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

Each $B_k(t-j)$ is incorporated into the model with its own coefficient, and the autoregressive model becomes:

$$\textit{Equation } 4.4 + \sum_{k} \sum_{j=0}^{l} \beta_{kj} B_k(t-j) \tag{4.6}$$

β_{kj} is the coefficient associated with $B_k(t-j)$. Using the log/linear form of Equation 4.6, we can estimate the general effect of a takedown using the estimates of β_{kj}. For each takedown, the fractional change in wickedness associated with the takedown during week j is $e^{\beta_{kj}} - 1$. This process can be repeated to give $e^{\sum_{j=0}^{l} \beta_{kj}} - 1$, which estimates the cumulative effect of the takedown over the time period l. If the MLE of any one of the β_{kj} is not statistically significant it is assumed to be 0. The statistical significance of the estimated coefficients provides a rigorous test of a takedown's effect.

We incorporated 12 different historical botnet takedowns into the extended model. We considered most major takedowns of botnets in the time span of our dataset that sent large amounts of spam. We allow i to vary from 0 (the week of the takedown) to $l = 6$ weeks. Beyond this time, we find no further statistically significant changes that can be attributed to the takedown, implying that the time horizon for the effect of a takedown is at most six weeks. In some cases, two botnet takedowns overlap the six-week windows, and we cannot separate the effect of the two takedowns.[8] When this occurs we include both the initial effect of the first takedown and the combined effect of the second takedown.

The results are given in Table 4.2, which shows that the global effectiveness of these botnet takedowns varies significantly. Some takedowns were effective in the short run (6 out of 12), but over the six-week window only three showed any persistent significant decrease in spam.

The table shows that two takedowns (Bredolab and Rustock) had a relatively

[8]An overlap results in two binary variables with the same value being included in the model (perfect collinearity), which would cause an ill-defined maximum likelihood calculation [298].

Table 4.2: Effect of 12 historical botnet takedowns in the model. The recorded dates are the first date in our data set after the intervention. In column two *Communications Disruption* is the severing of communication between bots and the command and control (C&C) infrastructure, *C&C Takeover* refers to when control of the C&C infrastructure is gained without physical access, *Seizure* refers to the physical confiscation of C&C infrastructure, and *Arrest* refers to the arrest of individuals. The percent change columns report the percent change in global wickedness in the first week after the takedown (column three) and six weeks later including the first week of the takedown (column 4).

Botnet takedown (Date)	**Takedown Method**	**Initial % Change**	**6 Week % Change**
McColo (11/11/08)	Communication Disruption [148]	-17.4	44.6
Mariposa (12/24/09)	C & C Takeover [56]	35.8	34.8
Waledac (3/5/10)	Communication Disruption [152]	Not significant	-3.5
Spamit.com (10/1/10)	Self Shutdown [285]	Not significant	6.1
Bredolab/Spamit.com (10/29/10)	Seizure and Arrest [83]	-11.8	-17.2
Rustock (3/19/11)	Seizure [139]	-20.2	-13.9
Coreflood/Rustock (4/16/11)	Communication Disruption [154]	-7.3	13.8
Kelihos (9/17/11)	Communication Disruptions [103]	6.4	31.6
Kelihos Variant (4/1/12)	Communication Disruption [72]	Not Significant	30.1
Hermes-Carberp (6/24/12)	Arrest [207]	21.4	9.0
Grum/Hermes-Carberp (7/22/12)	Communication Disruption [194]	-11.3	49.4
Virut (1/2213)	Communication Disruption [157]	-21.7	113.8

large long-term impact on spam in the six weeks following the takedown, while the third (Waledac) had a relatively minor impact. Both the Bredolab and Rustock takedowns involved physical seizure of infrastructure by law enforcement. Although this may not be directly related to the effectiveness of the takedowns, it is notable and is likely correlated with other external factors that have more lasting effect. Four takedowns that used communications disruption to shutdown the botnet reduced spam concentration in the short-term (i.e. McColo [148], Coreflood [154], Grum [194], and Virut [157]) are followed by long-term increases in wickedness. The rest of the takedowns, such as the self shutdown of spamit.com [285], seemed to have little positive impact either initially or in the long-term. These values provide evidence that other interventions were likely the main driver of the decline in overall spam volumes, not botnet takedowns. We note that the two most effective takedowns occurred at the end of era 1 and beginning of era 2 respectively, however, without more data we cannot to draw further conclusions about the relationship between takedown effectiveness and the era in which they occurred.

In the case of Mariposa, our results may reflect the historic details of the takedown. Shortly after the original takedown in December, during which control of command-and-control servers was obtained, attackers managed to regain control of the botnet and launched denial-of-service attacks against numerous ISPs [56], which could be related to the increased spamming activity.

4.5.2 Regional effects of botnet takedowns

Bots are not uniformly distributed geographically [196], suggesting that takedowns might have different effects throughout the world. To investigate this hypothesis we re-applied our modeling approach, but at the regional level. Rather than creating a single model for all ISPs globally, we constructed one model for each geographic

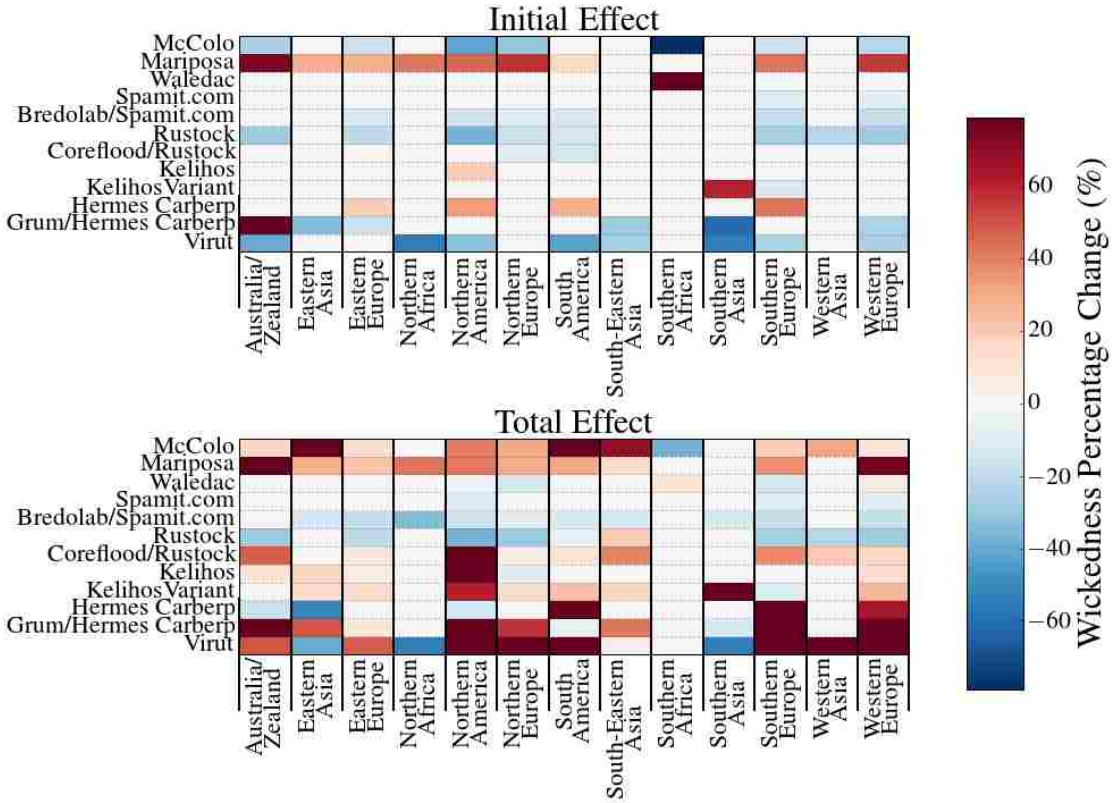

Figure 4.5: Regional effect of botnet takedowns. For each historical takedown studied the top panel shows the immediate effect by geographic region, and the bottom panel shows the effect after six weeks for the same geographic regions. The color shows the percent change in wickedness as indicated by the legend.

region defined in subsection 4.3.2, using only the ISPs in that region. We included regions that have at least two ISPs in our dataset to avoid over-fitting [298].

All takedowns showed varying effects for different regions (Figure 4.5). Some takedowns have effects regionally that resemble the global effect seen in Table 4.2, while others have differentiated behavior. For example, the McColo takedown initially appears successful, but in the long term wickedness increases across nearly all regions (blue colors, upper panel in Figure 4.5, and red colors, lower panel in Figure 4.5, respectively), similar to the global effect. In contrast, two of the takedowns

led to mixed effects throughout the world. Six weeks after the Hermes Carberp takedown, wickedness in Australia/New Zealand, Eastern Asia, and South-Eastern Asia decreased, but most other regions experienced increases. Similarly, six weeks following the Grum takedown, wickedness in South America had declined significantly, but the rest of the world experienced increases. These differentiated regional effects occur predominately in the second and third eras.

We can further analyze the effect of botnet takedowns on individual countries by constructing one model for each country, using the same procedure as we did for regions. Once again, we consider only countries with more than two ISPs in our dataset to avoid overfitting. Figure 4.6 shows the effect of various takedowns only on countries in Eastern Europe due to space constraints. We focus on Eastern Europe because it shows interesting variation among its countries. However, most other regions also showed significant variation.

Consistent with the earlier analyses, there are many countries for which a takedown initially has a positive effect, but where, in the long term, wickedness actually increases. One prominent example is the Czech Republic following the Bredolab/ Spamit.com takedowns, which did not experience a significant change in wickedness the week of the takedown, but wickedness nearly doubled after six weeks. Country-by-country there is little correspondence with the global takedown effect. For example, the McColo takedown initially reduced wickedness globally, but was followed by an increases in spam on both a global and regional level. However, at the country level the results are mixed, with Belarus benefiting from the takedown while Romania, Hungary, and Russia experience increases at 6 weeks.

These regional results raise the interesting possibility that botnets can migrate in response to takedowns. That is, by reducing the number of infected hosts in one region, a takedown creates incentives for botnets to find new vulnerable hosts, thus moving the problem elsewhere. More advanced modeling techniques, such as vector

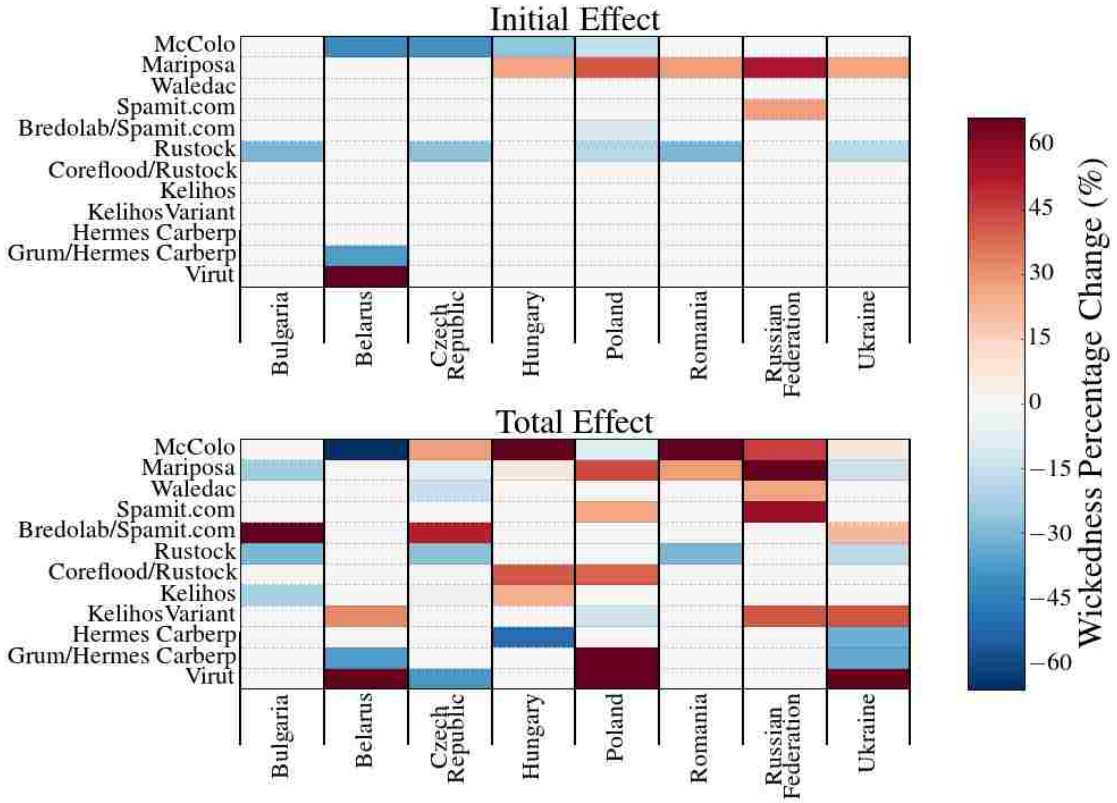

Figure 4.6: Country-specific effect of botnet takedowns in Eastern Europe. For each historical takedown studied, the top panel shows the immediate effect for each country, and the bottom panel shows the effect after six weeks for the same country. The color shows the percent change in wickedness as indicated by the legend.

autoregressive models [10], could shed light on this intriguing possibility.

4.6 Related Work

This chapter builds on the dataset of Van Eeten et al. [276], which investigated ISPs as control points for mitigating the spread of malware, using a comprehensive world-wide spam dataset. Here, we updated the dataset with 6 more years of data. The Van Eeten et al. analysis revealed that a country's development level is correlated with

spam volume, and it analyzed how public policy initiatives might reduce infections. We extend this work by developing a data-driven statistical model, which estimates the effect of different spam interventions and identifies temporal transitions in the dataset.

Other work locates infected hosts in IP address space. Moura et al. identified IP ranges with high concentrations of spam sending hosts [196]. Similarly Ramachandran et al. examined the network-level behavior of spammers, and showed that spam is concentrated in relatively small IP ranges [225]. Stone-Gross et al. studied ISPs with persistent malicious behavior [259], Chen et al. investigated malicious sources on the Internet over IPv4 [41], and Wilcox et al. studied the stability and availability of address space in spam and non spam networks [291]. Kokkodis and Faloutsos showed that spamming botnets have become more widely and thinly spread over IP space, a potential problem for filtering [144]. However, to our knowledge none of this work explores which topological features of the AS network correlate with infected hosts. Additionally, our model shows that previous regional concentrations of wickedness and malicious traffic correlate with future wickedness.

Collins et al. define *uncleanliness* as the probability that a host is vulnerable [52], while wickedness measures the concentration of active malicious hosts. They find that a network's past behavior is strongly correlated with its future behavior, which agrees with our finding that wickedness is autocorrelated.

Another related area proposes using economics to control malware and spam [135, 202, 169]. The idea of disrupting spammers' income by targeting the small number of banks that handle credit card payments [186, 136] may have helped reduce global spam levels. A related approach is the publication of infection rates of ISPs (measured by spam volumes) to provide incentives to control compromised customers in their networks [267].

There are few models of global malware dynamics. Venkataraman et al. model malicious activity as a decision tree over IP address space and infer the dynamics of the decision tree [279]. Their work focuses on IP address ranges rather than ISPs, but it reports some similar results as those observed in our model, for example, high variance in the data. Zhang et al. find that mismanagement of networks correlates with malicious behavior (measured using a quantity similar to our *wickedness*) in Autonomous Systems [308], but do not focus on how this behavior might evolve over time. Liu et al. use support vector machines trained on data from reputation blacklists to predict security incidents [176]. These predictions could be incorporated into our model to better predict some of the large changes in wickedness over time. A model of global malware dynamics was also proposed by Hofmeyr et al., which used an agent-based model to investigate the dynamics of malicious traffic flowing across the Internet at the AS level [118]. This model was significantly more abstract than ours, and did not incorporate actual data about spam, ISPs, demographic features, or intervention events such as takedowns.

Nadji et al. analyze botnet takedown efficacy [198], and other work considers raw measurements of spam volume [103]. Nadji et al. investigated three historical takedowns, performing post mortem analysis of each takedown's effectiveness, by measuring which malicious domains could still be resolved in the Domain Name Service (DNS). Contrary to our results, this work recommends DNS takedowns for a large fraction of current botnets. However, their results rely on relatively short time scales (two weeks), and it only considers the DNS, which may not be sufficient to identify rebounds once attackers establish new communication channels [100].

Mechanistic botnet models, e.g. [146, 301, 66, 42, 137], focus on specific infection mechanisms, while our model considers the security problem from a global perspective, with botnets being just one component. We find that most botnet takedowns have limited and transient impacts on global wickedness. This result agrees with

other research, which found that botnets are surprisingly resilient [303], and in many cases recover after a short time [198]. Other work has modeled malicious websites, noting the high variance of cybersecurity data, and investigates interventions through modeling [78].

Traffic filtering is an important intervention for reducing the number of infected hosts. There has been research into the effectiveness of various filtering techniques, e.g. [174, 133, 226], however this work focuses on the success of the filter itself and not whether the filter actually reduces the global distribution of infected hosts. Incorporating filtering interventions into our model is an area we plan to explore in future work.

4.7 Discussion

Data-driven models such as the one presented here can potentially yield interesting and important insights, which in turn can inform policy makers about the utility of interventions or even how to prepare vulnerable regions of the world before they are applied. However, there are several pitfalls that a statistical modeling approach needs to acknowledge.

First, the model is built around statistical correlations, but it ignores mechanisms, e.g. by what process does a country's development and an ISP's position in the ISP network influence wickedness? Second, statistical models such as ours cannot determine causality, so detailed understanding of the data is needed to attribute cause and effect. Third, high variance data can hide significant changes, and also make it appear that significant change has occurred when it has not. Modeling global data is a powerful tool to address this issue, but the modeling methodology must take into account the variance (e.g., averages can be misleading). We were careful to use appropriate methodology to avoid this pitfall. Finally, any conclusions drawn

from a statistical model depend on the quality of the data (although techniques do exist to help compensate for certain classes of data problems).

Any model is necessarily a simplification of reality. For example, our traffic model is simplistic given the complexity of the Internet. Future work could incorporate more realistic models, especially because our model shows that the traffic component is significant only during the third era. It could be that spam email was more likely to be used to spread infection during this era, whereas earlier it was used primarily for advertising, e.g. gray market pharmaceuticals. Similar to the traffic component, if other important features are identified, such as the type of service provided by an ISP, this information could easily be included in the model.

This chapter focused on spam itself, but spam data have also been used to estimate the numbers of infected PCs [312, 267, 132]. By applying our methodology to other measures of infection, it should be possible to develop models that provide insight into the dynamics and global distribution of these other types of infections. In general, we are interested in the distribution of all malicious behavior (or wickedness), regardless of its source. In some cases, the definition of wickedness could be expanded, e.g to include the relative value of hosts in different regions—an infected machine in the US may be more valuable than one in India.

Cybersecurity is often viewed as an arms race, which complicates the task of predicting the impact of today's interventions against tomorrow's attackers. At least, however, we should evaluate the likely effect of new methods before embracing large-scale deployments or policy directives that enforce certain interventions, and models such as the one described here are one way to approach this.

We have studied the impact of botnet takedowns in some detail, but there are other interventions that would also be interesting to explore. For example, the traffic model provides a way to analyze the effect of blacklisting offending ISPs or

different filtering strategies [116]. There is evidence that national and international initiatives against cybercrime can reduce wickedness [276]. Our model could, for example, be used to assess whether countries that are signatories to agreements such as the London Action Plan or Council of Europe's Convention on Cybercrime actually experience lower wickedness levels after ratifying the agreements. This could be studied by incorporating this information as an additional variable.

By looking for differential effects of takedowns geographically, we can identify *at risk* ISPs or countries, i.e. those that are likely to see little initial effect from the takedown but which could expect an increase in wickedness in the medium term. Our results to date have not identified any single factor that is consistently correlated (at a statistically significant level) with increased wickedness after a takedown. However, if we could identify at-risk countries and ISPs, they might make good candidates for targeted interventions, for example, ISPs on the periphery of the AS network which may have inadequate spam-fighting resources and lack automated methods to help customers clean up malware. Government interventions could focus on providing resources to those ISPs (or even countries), an approach that might prove more cost-effective than existing methods.

4.8 Summary

In this chapter we studied an abstract quantity called *wickedness* (concentration of spam sending hosts) and showed that it clusters regionally, correlating with national demographics and certain properties of the ISP graph. Through the use of statistical modeling combined with a large dataset, we studied some of the factors affecting spam, a large-scale security problem distributed around the world. Leveraging a long-term historical view of data produced interesting insights about the effectiveness of certain cybersecurity interventions. We found that takedowns are only marginally

effective in many cases, and in fact may be harmful to certain countries and ISPs.

Our model could serve as a starting point to predict future wickedness and test the likely effect of new interventions, both for spam and other similar problems. Our ultimate goal is to provide researchers and policy makers objective means to test intervention strategies and decide how best to mitigate global wickedness.

Chapter 5

Modeling Malware Spread and the Effect of Interventions[1]

The previous chapter demonstrated how longitudinal data about security phenomena and defenses can help measure the impact of interventions. However, such data are not always available. This is the case when examine malicious web pages presented in search results. Comprehensive data on malicious webpages is likely held by search companies such as Google and Microsoft, however, it is difficult for researchers to collect. Luckily, an approach which uses abstract models built on reasonable assumptions can still reveal insights about the effect of interventions. In this chapter we present such an abstract model.

[1]Material in this chapter was previously published as "Beyond the Blacklist: Modeling Malware Spread and the Effect of Interventions," which appeared in the *Proceedings of the 2012 Workshop on New Security Paradigms*. The original idea for this research was conceived by Tyler Moore. I conceived and analyzed the underlying model, and my co-authors aided in the preparation of the final written presentation.

5.1 Introduction

The network worms that caused havoc ten years ago, such as Code Red, actively spread by 'pushing' themselves onto vulnerable systems through automated scanning. In contrast, a major problem today is computer infections that propagate via a 'pull'-based mechanism. For example, in a drive-by download, an attacker infects a victim computer's web browser without direct interaction [222, 221]. In this scenario, the attacker first compromises an otherwise benign web server, injecting executable code into its web pages, and then waits for users to visit the infected website and acquire the infection. Because many users arrive at websites through search, search engines are a crucial battleground over the distribution of malware.

Search providers have an incentive to defend against such attacks because they degrade search results. A typical approach is that taken by Google, which attempts to detect and blacklist websites that host malicious content [104]. Blacklisting can take the form of displaying a warning message via a client side browser plugin to discourage users from visiting a website, or outright removal from the search results. Blacklisting can be used to combat many types of malicious content, which is important in a web environment where new attacks are developed frequently. However, because blacklisting can dramatically reduce visits to websites, search engines are careful to avoid false positives (i.e., flagging an uninfected website as infected). Such caution can delay responses, which in turn may raise infection rates.

In this chapter, we devise a concise Markov model to study how web infections spread through large populations of websites, and explore how infections might be contained with blacklisting. We also propose a generalization of blacklisting called *depreferencing*, where a search engine reduces a website's ranking in search results in proportion to the engine's certainty that the website is infected. Depreferencing can be more tolerant of false positives than a binary response such as blacklisting

because the scale of the intervention can be adjusted to specific levels of false positives. Depreferencing provides a controllable *depreferencing parameter*, σ, that can be tuned to achieve specific reductions in infections or false positives. We derive exact analytic expressions that relate the depreferencing parameter, σ, to infection rates and traffic loss due to false positives. We also identify critical points for the model parameter values that govern the trade-off between infection and traffic loss.

We believe that modeling is particularly well-suited to the task of examining techniques for controlling malware spread over the web. First, it allows us to examine unconventional interventions, such as depreferencing, at low cost. Given the relatively grim status quo in web security,[2] more radical countermeasures deserve consideration, and modeling offers a good way to assess the impact of new strategies without the expense and commitment of an actual implementation.

Second, modeling can deal with the extreme dynamics of the web better than empirical exploration alone. Our analysis shows that the heavy-tailed distribution of website popularity leads to high variance in outcomes. It is often possible for online services such as search engines to perform experiments by rolling out improvements to subsets of clients. However, high sample dependence makes it extremely difficult to conduct reliable comparative assessments of the benefits of different interventions, especially with a limited number of empirical measurements. For example, the number of known malicious IP's can vary wildly over time, as shown in Figure 5.1. Extremely high values could conceivably be the result of a very popular website becoming infected. We show that this variance can obscure even large improvements in infection and recovery. With the modeling approach, we can easily run many simulations, and more reliably estimate the comparative impacts of different intervention strategies.

[2] A 2011 report found that 84% of websites were vulnerable to attack for more than 30 days of the 2010 calendar year [289].

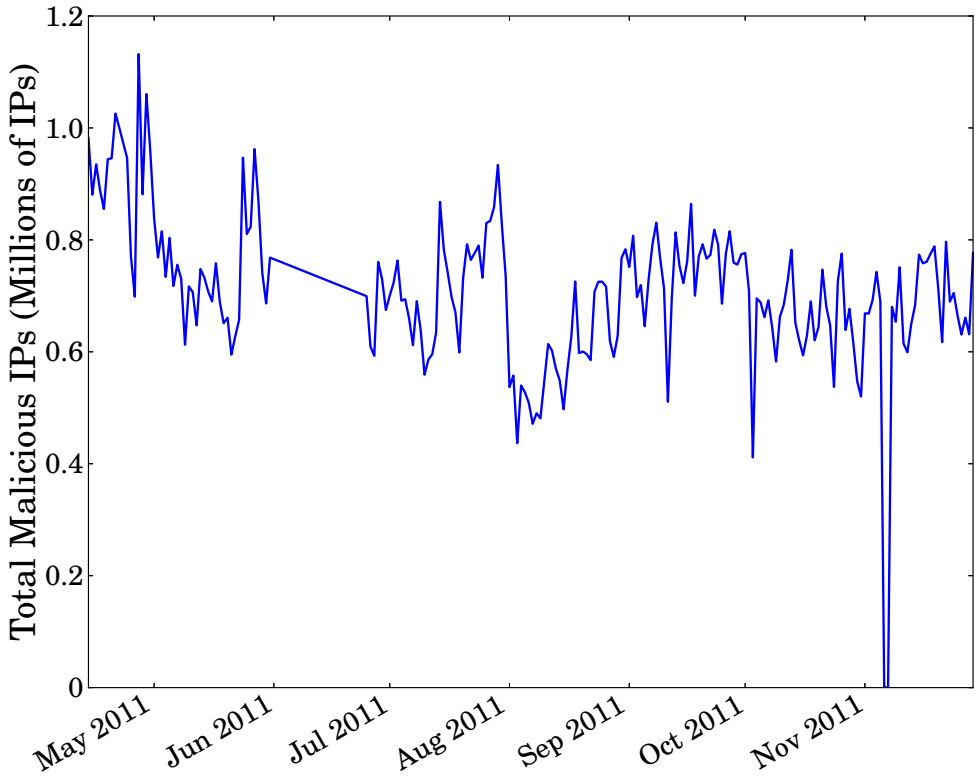

Figure 5.1: Variation in malicious IP addresses over time; from the Internet Storm Center (`http://isc.sans.org`)

Finally, modeling lets us examine the impact of interventions across many stakeholders and identify tensions that may arise. For instance, improved security for search operators and consumers may be achieved in part at the expense of increased risk of incorrect blacklisting for website operators. Modeling allows us to more precisely quantify these trade-offs.

5.2 Modeling Infections

We model a population of servers that is under attack from malicious agents, as depicted in Figure 5.2. We do not model specific types of infections, assuming that an infection is any event that compromises a website such that it could be used to spread malware to users. Once infected, a server recovers when an administrator notices the infection and clears it. In this chapter we explore the impact of search provider interventions and so are only interested in clients that connect to servers via referrals from a search provider. Hence, in our model, client exposure to infection is driven solely by website popularity as determined by the search provider. In an attempt to improve search results, the search provider monitors websites to determine whether they are infected and may incorrectly identify uninfected websites as infected. We assume that an administrator clears false identifications of infection at the same rate as real infections.

Our model includes a population of n websites[3], each with a popularity, ω_i, drawn at random from a specified distribution. ω_i represents the total number of visits a website receives. The key outcome we are interested in measuring is *client exposure*, which is directly proportional to the expected number of visits that infected websites receive. At any time, a website is in one of three possible states: infected, uninfected, or falsely infected (i.e. classified by the search provider as infected when it is actually not). Each server transitions between these states at discrete time steps, according to the Markov chain depicted in Figure 5.3. The key parameters are: ρ, the probability of a website becoming infected; γ, the probability of recovering from an infection; and f, the probability of falsely being classified as infected.

We make the simplifying assumption that the probabilities ρ, γ, and f are constant across the population of servers and time invariant.[4] Unfortunately, data on

[3]We use the terms website and web server, or simply server, interchangeably.

[4]In Section 5.5.5 we briefly explore the impact of relaxing this assumption.

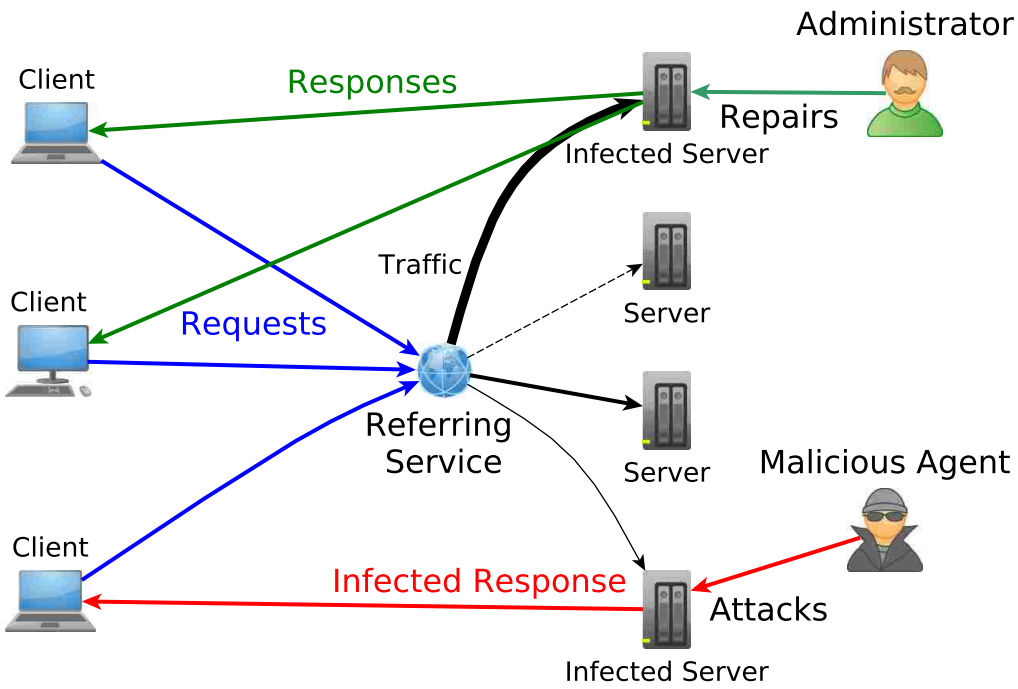

Figure 5.2: Server and client infections via search engine referral.

the exact distributions of these parameters are not readily available and often contradictory. For example, there are no data supporting a systematic relationship between a website's popularity and its susceptibility to infection. Although Moore et al. [193] found that more popular web search terms are less likely to include infected websites in their results, it is possible that more popular sites are higher priority targets for exploitation and therefore more likely to be infected. Because in both cases the effects are likely small we argue that assuming constant probabilities is reasonable.

Our model is discrete time; an alternative approach is to model the population of servers using differential equations. In the case of large n, the steady state distribution of infection probability would be exactly the infection rates in a differential

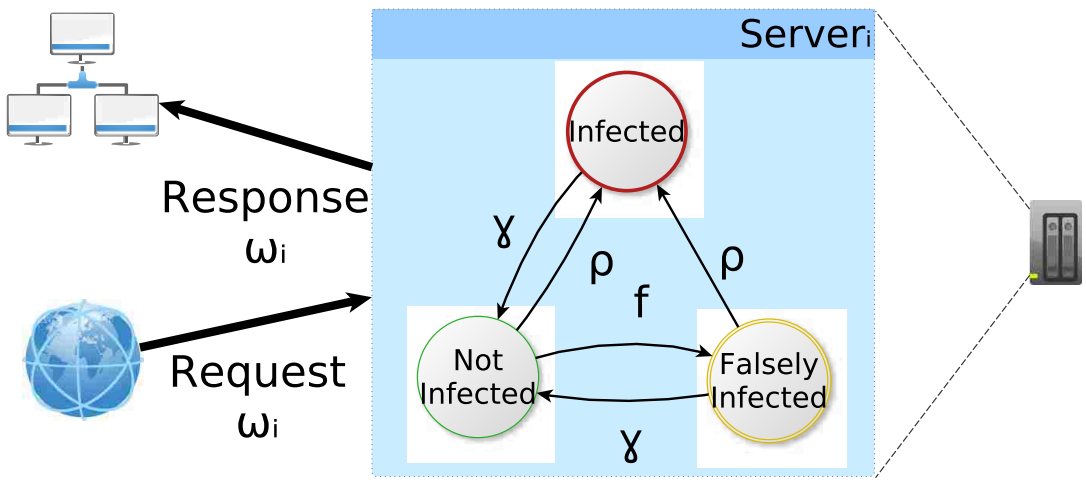

Figure 5.3: Model of website infections and client exposure.

equation model[313]. We use a discrete-time model instead because it allows us to easily incorporate time-dependent phenomena (such as interventions) and distributions of values (such as traffic), and it is simpler to explore transient effects.

5.3 Modeling Interventions

We model two forms of intervention: *blacklisting*, which is currently used by search engines, and a hypothetical approach called *depreferencing*, which offers a way to adjust intervention parameters to specifically control the trade-off between infections and traffic loss due to false positives.

5.3.1 Blacklisting

A common approach taken by search engines that detect a compromised website is to inform the user in the search results or through a client side application before the user has a chance to visit the website, and then to produce further warnings if the user persists in attempting to visit the website. Although research has shown that users will often disregard warnings, such as certificate warnings [255, 263] and phishing attack warnings [170, 79], these studies do not investigate the typical multi-step warning process presented in search results. Search result warnings are difficult to circumvent, and we believe that users are likely to simply choose an alternative result (provided they are not trying to reach a specific site). Additionally, certificate warnings are commonplace due to badly designed APIs of SSL implementations[97], which may lead to the high level of circumvention.

For example, if Google identifies a site as malicious, the text "This site may harm your computer", will appear under the link in the search results. Attempting to proceed will take the user to a warning page with a small text URL that must be copied and pasted (without instruction) into the browser navigation bar to proceed. If the user persists, client side browser tools will present yet another warning page with a small link to "proceed anyway" (Chrome Browser) or "ignore this warning" (Firefox Browser). This is a far cry from a simple click through dialog.

Given that the goal of blacklisting is to prevent a website from receiving all or nearly all of its search traffic, minimizing false positives is essential. For example, Rajab et al. [224] claim that Google's Safe Browsing infrastructure "generates negligible false positives", and Google themselves state they "... strive for high quality and have had only a handful of false positives" [208]. While the specific amount of resources dedicated to correctly identifying malicious websites remains unknown outside of search providers, Google states that it "... invests heavily in the Safe Browsing

team." [208] Moreover, the fact that resources any resources are being dedicated to this problem indicates that Google believes malicious websites to be detrimental to a user experience.

We assume that blacklisting takes a fixed number of time steps to detect a compromised website and blacklist it. We refer to this as the *detection delay*, denoted β. A website infected at time t will be blacklisted at time $t + \beta$. Once blacklisted, the traffic to that the website is set to zero, i.e. $\omega_i = 0$. Formally, if a website, i, is infected at time x, its traffic, $\hat{\omega}_i$, at time $t > x$ is

$$\hat{\omega}_i = \begin{cases} \omega_i, & \text{if } t - x < \beta \\ 0, & \text{if } t - x \geq \beta \end{cases} \tag{5.1}$$

The time period β captures the notion that it will take a search engine a certain amount of time to determine that a website is compromised with high certainty (negligible false positives). Thus β accounts for how frequently the website is crawled, how much computational time is required to confirm the infection, how much the search engine is willing to invest in malware detection, and other possibilities, such as giving the compromised site a certain grace period to clean up the infection. We do not have good information on the specific costs associated with β. However, the specifics are irrelevant to the outcomes of our model, and search engines can use their own data to evaluate each action that contributes to β.

In the model, we assume that immediately after a website recovers, its popularity is restored to its previous value. That is, once a website has been cleaned, the administrator informs the search engine and the blacklisting is removed without delay. In reality, there would be a small delay before the blacklisting is removed. For example, when an administrator requests Google to run an automated test for malware, it will take at most a few hours to complete, and up to 24 hours for the

malware warning to disappear from all search results [216]. Because the time period is small and constant, we can exclude it from our model without significantly changing the results.

5.3.2 Depreferencing

We explore a generalized hypothetical intervention, called *depreferencing*, which, to the best of our knowledge, is not actually implemented by any existing search engine. The idea is that when a search engine detects a possibility of infection in a website, it reduces the traffic that website receives. This could be implemented by reducing the rank of that website in the search results, or probabilistically providing warnings to users. Because the response does not block all traffic to the website, but rather reduces the volume of traffic, the detection process can tolerate false positives, allowing the search engine to react more rapidly and aggressively. Search providers could use potentially coarser and less precise detectors to crawl websites more frequently, requiring significantly less computation time to classify websites as infected.

We model this intervention by reducing the popularity of a website by a fixed percentage every time step after it is discovered that the website is infected. If a website is infected at time x, an infected website's traffic at time $t > x$ is

$$\hat{\omega}_i = \begin{cases} \omega_i, & \text{if } t - x < \beta \\ \sigma^{(t-x)-\beta+1}\omega_i, & \text{if } t - x \geq \beta \end{cases} \tag{5.2}$$

where $0 \leq \sigma \leq 1$ is the *depreferencing* /parameter, which controls the strength of the response. Note that Equation 5.1 is equivalent to Equation 5.2 when $\sigma = 0$. We believe that adjusting search results is a plausible response that would be easy to implement. For example, a search engine like Google could simply reduce the page

ranks of infected websites, which would directly affect their popularity in search results. Similarly to blacklisting, we assume that when a website recovers from an infection, its popularity is immediately restored to its original value. Because depreferencing is a less drastic response, search engines might be able to reduce the detection delay β if they were to adopt this intervention.

Equation 5.2 is one of an even more general class of methods for combating exposure to infection. We could define a general $g(\omega_i, x)$, such that g is monotonically decreasing in time. For example g could be a linear or logistic function. We choose an exponential decline as it seems a natural fit for our application. Investigation into other forms appropriate for other applications is left for future work.

As a consequence of the potentially more rapid, and hence, imprecise detection of compromised sites, our model includes a constant probability f that an *uninfected*/website is classified as compromised and has its rank reduced. This is in contrast to the blacklisting approach, where we assume there are zero false positives. For depreferencing, we assume that websites that are incorrectly classified as compromised recover at the same rate, γ, as compromised websites. In other words, the process of recovery is the same whether a website is actually infected or not. This requires that the administrator realize that the website is infected (for example, users of Google's Webmaster Tools are notified when their sites are infected) and that appropriate steps are taken to correct the problem.

We do not model false negatives, i.e. infected websites that go undetected, because our model studies the effect of interventions on client infection rates, and we assume that in both blacklisting and depreferencing the detection process has similar levels of false negatives. Hence, the false negative rates should not affect comparison of the outcomes. From a practical perspective, data on false negatives are rare or nonexistent because they are extremely difficult to gather. We leave the analysis of false negatives to future work.

5.4 Analysis

This section analyzes the mathematical properties of the model described in the previous section. First we describe the steady state values of the Markov chain shown in Figure 5.3. Second, we analyze the first and second moments of the random variables that define the traffic loss and the number of clients exposed to infection. We then provide expressions that relate the intervention parameters to the infection exposure and traffic loss, and identify critical control points.

5.4.1 Steady State Distribution

Let the state of a server i in the Markov chain in Figure 5.3 be the random variable $S_i \in \{I, N, F\}$, where I denotes infection, N denotes no infection and F denotes a false positive infection. It is easy to see that the Markov chain is ergodic except for some degenerate cases such as $f = 1, \gamma = 1, \rho = 0$. However, such cases are unlikely to occur in the real world.

Because our Markov chain is ergodic it is guaranteed to converge to a unique stationary distribution, which is given by

$$Pr[S_i = I] = \frac{\rho}{\rho + \gamma} \tag{5.3}$$

$$Pr[S_i = N] = \frac{\gamma}{(f + \rho + \gamma)} \tag{5.4}$$

$$Pr[S_i = F] = \frac{f\gamma}{(\gamma + \rho)(f + \gamma + \rho)} \tag{5.5}$$

Moreover, because this is a finite time-homogeneous ergodic Markov chain, it will have a short mixing time. Hence we focus on the steady-state in the remainder of the analysis.

5.4.2 Client Exposure and Website Loss

The probability that a website becomes infected at a time $t-x$ and remains infected until time t depends on the probability that the website was not infected at time $t-(x+1)$, became infected at time $t-x$, and remained infected for the next x timesteps. More formally, let I_x denote the event that a server i has been in a state of infection for *exactly* x time steps. Then

$$Pr[S_i = I_x] = \rho(1 - Pr[S_i = I])(1-\gamma)^x \tag{5.6}$$

Observe that the events $S_i = I_x$ and $S_i = I_{x'}$, with $x \neq x'$, are mutually exclusive, e.g. a server cannot be infected for exactly 5 and exactly 6 time steps.

Next we derive an expression for the random variable $X_i(\beta, \sigma)$, which describes the number of clients exposed to infection from a website i, when the search provider implements an intervention controlled by the parameters β and σ. Recall that β is the detection delay for infection identification and σ is the depreferencing parameter, i.e., the strength of the response. The expectation of exposure to infection from website i is then

$$\begin{aligned}\mathbb{E}[X_i(\beta,\sigma)] =& \sum_{x=0}^{\beta-1} \omega_i \frac{\rho\gamma(1-\gamma)^x}{\rho+\gamma} + \\ & \sum_{x=\beta}^{\infty} \omega_i \sigma^{x-\beta+1} \frac{\rho\gamma(1-\gamma)^x}{\rho+\gamma} \\ =& \frac{\omega_i \rho\gamma}{\rho+\gamma}\left[\frac{1-(1-\gamma)^\beta}{\gamma} + \frac{\sigma(1-\gamma)^\beta}{1-(\sigma(1-\gamma))}\right]\end{aligned} \tag{5.7}$$

The above expression simplifies to $\omega_i Pr[S_i = I_i]$ when no intervention is taken, which would correspond to $\beta = \infty$ or $\sigma = 1$.[5]

[5]Because this expression has two parts, namely infection spread pre and post interven-

The other important random variable we are interested in is $L_i(\beta, \sigma)$, which represents the traffic lost by a website i as a consequence of false positives. Following a similar analysis to the earlier one for client exposure, if F_x denotes being in the false positive state for x time steps, we have

$$Pr[S_i = F_x] = fPr[S_i = U](1 - (\gamma + \rho))^x \tag{5.8}$$

The lost traffic at a specific time will be $\omega_i - \hat{\omega}_i$. Substituting for $\hat{\omega}_i$ as given by Equation 5.2, the expected traffic loss is

$$\begin{aligned} &\mathbb{E}[L_i(\beta, \sigma)] = \\ &\frac{\omega_i f \gamma (1 - (\rho + \gamma))^\beta}{f + \gamma + \rho} \left[\frac{1}{\gamma + \rho} - \frac{\sigma}{1 - \sigma(1 - (\gamma + \rho))} \right] \end{aligned} \tag{5.9}$$

We can then define the *infection exposure*, which is the fraction of traffic exposed to infection from all websites, as

$$X(\beta, \sigma) = \frac{\sum_i^n X_i}{\sum_i^n \omega_i} \tag{5.10}$$

and the overall *traffic loss* due to false positives as

$$L(\beta, \sigma) = \frac{\sum_i^n L_i}{\sum_i^n \omega_i} \tag{5.11}$$

Using linearity of expectation, the expressions for $\mathbb{E}[X(\beta, \sigma)]$ and $\mathbb{E}[L(\beta, \sigma)]$ are simply those in Equation 5.6 and Equation 5.8 respectively, while omitting ω_i, specifically:

---

tion, we could easily include two recovery rates γ_{pre} and γ_{post} to model the fact that recovery likely occurs more quickly after an intervention is taken. This does not significantly effect our results here or in experiments.

$$\mathbb{E}[X(\beta,\sigma)] = \frac{\rho\gamma}{\rho+\gamma}\left[\frac{1-(1-\gamma)^{\beta}}{\gamma}+\frac{\sigma(1-\gamma)^{\beta}}{1-(\sigma(1-\gamma))}\right] \tag{5.12}$$

$$\mathbb{E}[L(\beta,\sigma)] = \frac{f\gamma(1-(\rho+\gamma))^{\beta}}{f+\gamma+\rho}\left[\frac{1}{\gamma+\rho}-\frac{\sigma}{1-\sigma(1-(\gamma+\rho))}\right] \tag{5.13}$$

We note that both of the infection exposure and the traffic loss are independent of the distribution from which the ω_i's are drawn, or how many servers there are.

The effectiveness of the depreferencing parameter, σ, and the detection delay, β, in the control strategy for $\mathbb{E}[X(\beta,\sigma)]$, depends only on the recovery rate γ. If $\gamma \approx 1$ (a fast recovery rate), then $\mathbb{E}[X(\beta,\sigma)] \approx Pr[S_i = I]$, i.e. the expected infection exposure is approximately the probability of a single server being in the infected state. Only when websites are slow to react to infections are interventions which alter traffic likely to have significant impact.

Conversely, ρ and γ both affect $\mathbb{E}[L(\beta,\sigma)]$. In particular, a decrease in a particular websites infection rate ρ or the recovery rate γ will cause an increase in loss due to false positives for a fixed false positive rate f. Intuitively, a website that is unlikely to be in the infected state is more vulnerable to being *falsely infected*.

We now determine the variance in $X(\beta,\sigma)$ and $L(\beta,\sigma)$. Because each of the X_i's is independent and the sum of the traffic is a constant,

$$Var[\frac{\sum_i^n X_i}{\sum_i^n \omega_i}] = \frac{\sum_i^n Var[X_i]}{\left(\sum_i^n \omega_i\right)^2} \tag{5.14}$$

Additionally, variance can be defined as

$Var[X_i(\beta,\sigma)] = \mathbb{E}[X_i(\beta,\sigma)^2] - \mathbb{E}[X_i(\beta,\sigma)]^2$. Using these two facts and some simple algebra we have:

$$Var[X(\beta,\sigma)] = \left(\mathbb{E}[X(\beta,\sigma^2)] - \mathbb{E}[X(\beta,\sigma)]^2\right) \frac{\sum_{i=1}^{n} \omega_i^2}{(\sum_{i=1}^{n} \omega_i)^2} \tag{5.15}$$

If the ω_i's are drawn from a distribution with finite variance and expectation and n is large, then we can apply the central limit theorem to Equation 5.15 to rewrite it in terms of the distribution of ω_i's

$$\begin{aligned} &Var[X(\beta,\sigma)] = \\ &\left(\mathbb{E}[X(\beta,\sigma^2)] - \mathbb{E}[X(\beta,\sigma)^2]\right) \left(\frac{Var[\omega_i] + \mathbb{E}[\omega_i]^2}{n\mathbb{E}[\omega_i]^2}\right). \end{aligned} \tag{5.16}$$

Observe that Equation 5.16 is monotonically decreasing in the number of servers n. So as the population of websites increases we expect the variance in the fraction of traffic exposed to infection to go to 0.

It is almost certain, however, that the distribution of ω_i for real webservers is heavy-tailed and does not have finite variance or finite expectation [3, 48, 187]. In the case of a heavy-tailed or power-law distribution of ω_i, the variance $Var[X]$ does not converge to a single value for large n, but to a distribution of values. Furthermore, because the sum of power-law i.i.d. random variables exhibits heavy tailed behavior [283] [99], the distribution of $Var[X(\beta,\sigma)]$ will also exhibit heavy tailed behavior.

The sum of power law distributed variables can be approximated by the maximum over the variables [306], which means that the last fraction in Equation 5.15 can be approximated as 1 for particularly heavy tailed distributions and large n, i.e.

$$\frac{\sum_{i=1}^{n} \omega_i^2}{(\sum_{i=1}^{n} \omega_i)^2} \to 1 \tag{5.17}$$

If we take this as an upper bound, we see that improving either σ or β to lower infection will also lower the variance in the infection exposure rate. Depending on the value of the exponent in the distribution of traffic, $Var[X(\beta, \sigma)]$ may not have finite variance or expectation. As we discuss later, this is important because it implies that empirical studies of infection exposure (or traffic loss) are likely to be highly sample dependent, and that even significant changes to the variables like ρ and γ can be hard to discern.

A similar analysis yields slightly different results for traffic loss:

$$\begin{aligned} &Var[L(\beta, \sigma)] = \\ &\left(2\mathbb{E}[L(\beta, \sigma)] - \mathbb{E}[L(\beta, \sigma^2)] - \mathbb{E}[L(\beta, \sigma)]^2\right) \frac{\sum_{i=1}^{n} \omega_i^2}{(\sum_{i=1}^{n} \omega_i)} \end{aligned} \tag{5.18}$$

5.4.3 Critical Values

In general, changing parameter values from one set, (β, σ), to another, (β', σ'), will result in a change in infection exposure, i.e., $\mathbb{E}[X(\beta, \sigma)] \neq \mathbb{E}[X(\beta', \sigma')]$. However, there could be some settings of β' and σ', such that the outcome will not change, i.e., $\mathbb{E}[X(\beta, \sigma)] = \mathbb{E}[X(\beta', \sigma')]$. We call these settings, or transition points, the *critical* values for the parameters.

The critical value, σ_X, for the depreferencing parameter is the most important, because we expect that search providers will have more control over σ than β. For example, a new detection algorithm may require a different β'; the search provider could then use the critical value of σ_X to ensure that the infection exposure did not change.

To derive the critical value for the infection exposure, we first calculate an expression for the precise value of σ needed to achieve a particular infection exposure

rate

$\mathbb{E}[X(\beta, \sigma)] = \xi$, as

$$\sigma = \frac{\frac{(\rho+\gamma)\xi}{\rho\gamma} - \frac{1-(1-\gamma)^{\beta}}{\gamma}}{(1-\gamma)\left[\frac{(\rho+\gamma)\xi}{\rho\gamma} - \frac{1-(1-\gamma)^{\beta}}{\gamma}\right] + (1-\gamma)^{\beta}} \tag{5.19}$$

We can then derive the critical value for the infection exposure by substituting $\mathbb{E}[X(\beta', \sigma')]$ for ξ in Equation 5.19, which gives

$$\sigma_X = \frac{a}{\gamma + a(1-\gamma)} \tag{5.20}$$

where a is defined as

$$a = 1 - (1-\gamma)^{\beta-\beta'} + \frac{\sigma\gamma(1-\gamma)^{\beta-\beta'}}{1-\sigma(1-\gamma)} \tag{5.21}$$

Equation 5.21 shows the critical value needed to ensure the infection exposure does not change when β changes. An alternative goal might be to ensure that the traffic loss due to false positives does not change with a new value for β, i.e. $\mathbb{E}[L(\beta', \sigma')] = \mathbb{E}[L(\beta, \sigma)]$. This will be given by another critical value, σ_L. Once again, we first derive an expression for the precise value of σ needed to attain a particular expected traffic loss fraction $\mathbb{E}[L(\beta, \sigma)] = \lambda$,

$$\sigma = \frac{\frac{1}{\rho+\gamma} - \frac{\lambda(f+\gamma+\rho)}{f\gamma(1-(\rho+\gamma))^{\beta}}}{1 + (1-(\rho+\gamma))\left[\frac{1}{\rho+\gamma} - \frac{\lambda(f+\gamma+\rho)}{f\gamma(1-(\rho+\gamma))^{\beta}}\right]} \tag{5.22}$$

Setting $\mathbb{E}[L(\beta', \sigma')] = \lambda$ in Equation 5.22, we get

$$\sigma_L = \frac{b}{1 + b(1-\gamma-\rho)} \tag{5.23}$$

where b is defined as

$$b = \frac{1}{\gamma + \rho} - (1 - \gamma - \rho)^{\beta - \beta'} \left[\frac{1}{\gamma + \rho} - \frac{\sigma}{1 - \sigma(1 - \gamma - \rho)} \right] \tag{5.24}$$

As can be seen from Equation 5.23, the critical value for the traffic loss is independent of the false positive rate f.

Using Equation 5.20 and Equation 5.23 in combination, a search provider has the ability to decide how to adjust σ to balance an increase in the traffic loss against an increase in infection exposure.

5.5 Experimental Results

To verify the results derived in Section 5.4 we used a Monte Carlo simulation of the model described in Section 5.2. Unless otherwise noted, we used the following parameter settings for all experiments: $\rho = 0.01$, $\gamma = 0.1$, and $n = 1000$. Although we believe that these parameter settings are plausible, our goal is not to provide a precise match with real-world outcomes, but rather to investigate more general consequences of features such as variance and the comparative efficacy of interventions. For each experiment, we conducted 1000 runs, and each run was 75 time steps. This length is sufficient for the model to reach a steady state.

We examine two different distributions throughout the experiments: uniform, with $\omega_i \propto Uniform(0, 1)$, and power law with $\omega_i \propto x^{\alpha}$ with $\alpha = -1.4$. Although these two distributions are likely not precisely representative of the real world, they are useful in that they represent two possible extremes of variance (finite and undefined).

In reality, the distribution is likely heavy-tailed, possibly a power-law [3, 48, 187]. We found that a power-law with an exponent of $\alpha = -1.4$ provides a good fit with

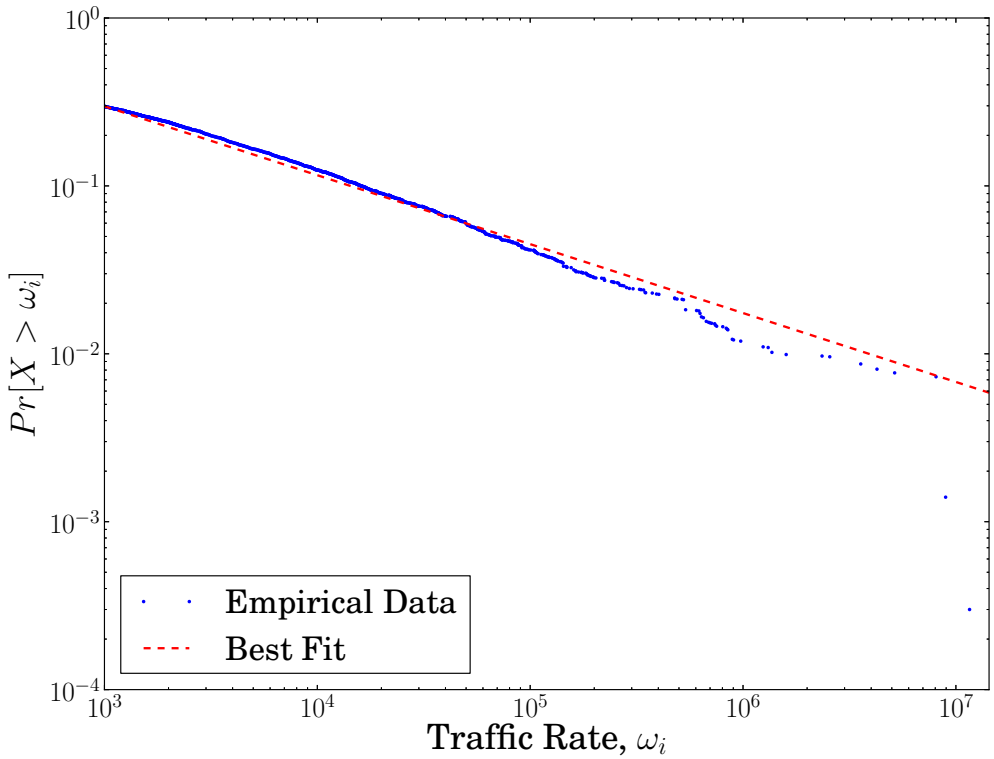

Figure 5.4: Empirically observed website traffic follows a power-law distribution with $\alpha = -1.4$.

empirical data on website popularity, as can be seen in Figure 5.4. We calculated the exponent for a random sample of 10,000 websites listed in the top 1 million websites according to the web-analytics firm Alexa, using estimates for the daily number of visits obtained by querying the Alexa Web Information Services API.[6]

[6] `http://aws.amazon.com/awis/`

5.5.1 Popularity Distribution

According to the analysis in Section 5.4, distributions of website popularity with undefined variance will result in large fluctuations in client exposure to infection and will be highly dependent on the sample of servers chosen. This is confirmed in our experiments, as can be seen in Figure 5.5 and Figure 5.6. The uniform distribution of website popularity results in low variance in client exposure (Figure 5.5), whereas the power law website popularity results in very high variance, both in a single run of the model and among different runs (Figure 5.6).[7] For both popularity distributions, the experimental average of the runs rapidly converges to the expected steady-state value for X (0.091), although power-law distributions can yield X values as high as 0.96 in individual runs, an order of magnitude higher than the expected value.

Figure 5.7 and Figure 5.8 show the variation in individual runs more clearly. Figure 5.7 shows three different runs of the simulation with the same parameters, $\rho = 0.01, \gamma = 0.1$. There are large jumps in client exposure to infection that occur when the more popular websites get infected, followed by plateaus before those websites recover, and then abrupt drops after recovery. Figure 5.8 shows two runs of the model with different infection and recovery rate parameters. Strikingly, the run with the infection rate cut in half and the recovery rate doubled, seems to exhibit worse infection behavior. This clearly illustrates why it might be difficult to determine whether web security improvements are effective. The high variance in the runs illustrates the importance of modeling, as running experiments in the real world could require many trials over long periods of time to reach conclusions with any confidence.

We also tested distributions other than uniform and power-law and confirmed the

[7]Because the variance is undefined in general for a power-law, we substitute the run sample values of the ω_i's into Equation 5.15 to compute the theoretical variance shown in Figure 5.6.

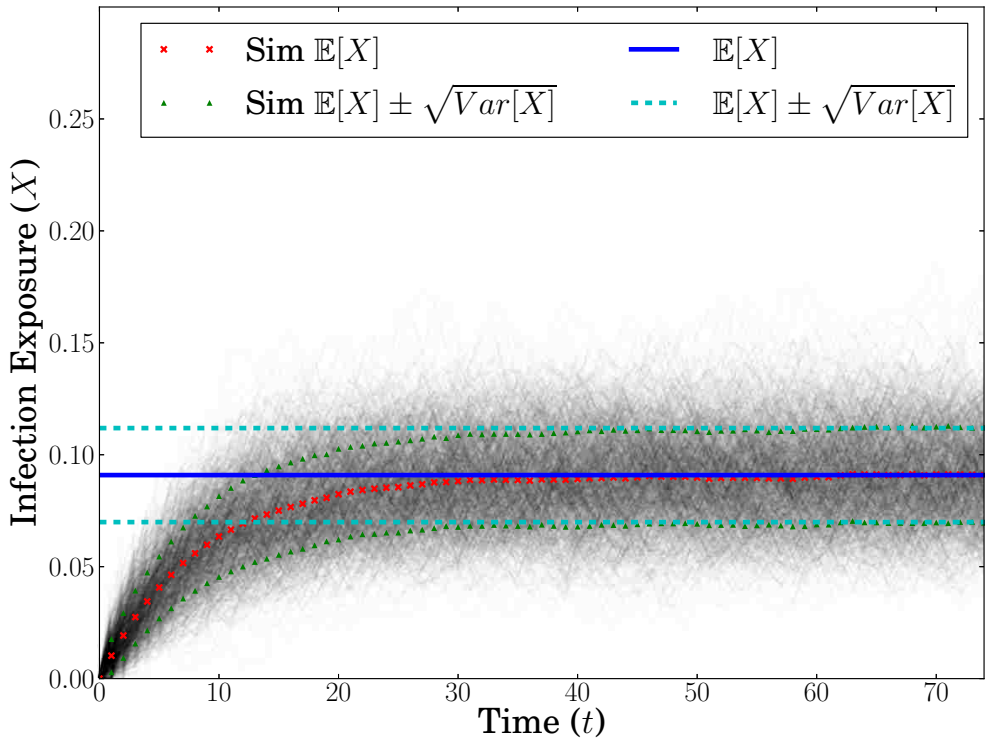

Figure 5.5: Variation in client exposure to infections over time when websites are selected uniformly at random. Individual runs are light grey, *Sim X* indicates the result of the simulation. Here $n = 250$ to illustrate the effects of small sample sizes. $\rho = 0.01$, $\gamma = 0.1$.

theoretical prediction that distributions with finite variance produce low variance in the measured outcome, whereas those with undefined variance produce high variance in the measured outcome (results not shown).

5.5.2 Interventions

Figure 5.9 and Figure 5.10 demonstrate the effect of varying the detection delay, β, on the steady state client exposure rate. For both uniform and power-law popularity

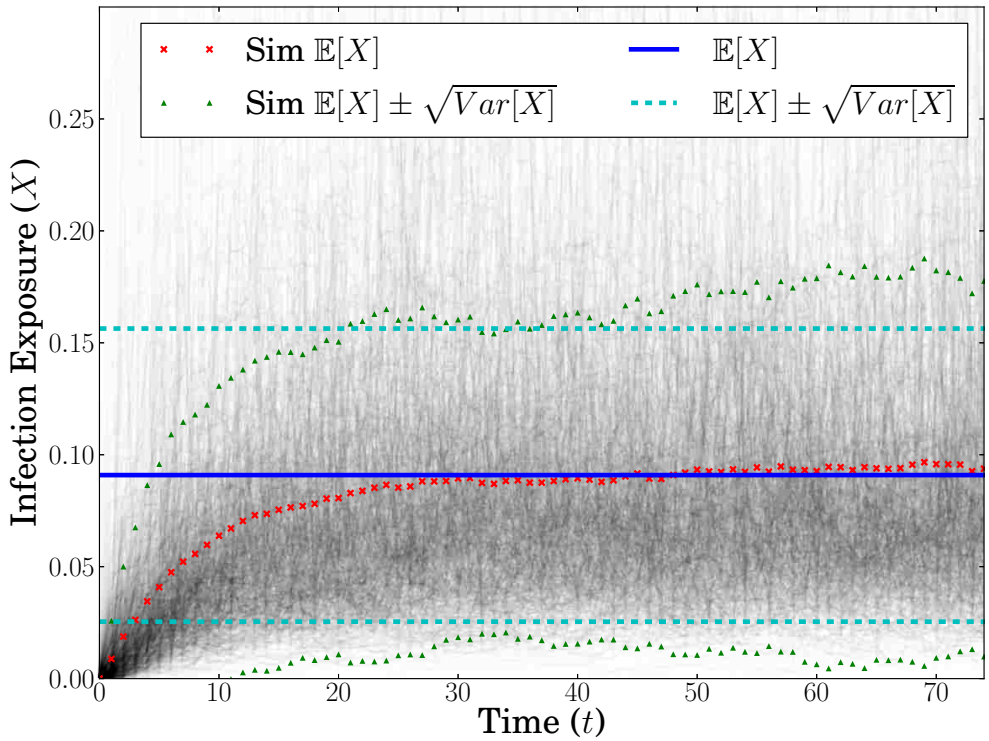

Figure 5.6: Variation in client exposure to infections over time when websites are selected from a powerlaw distribution. Individual runs are light grey, *Sim X* indicates the result of the simulation. Here $n = 250$ to illustrate the effects of small sample sizes. $\rho = 0.01$, $\gamma = 0.1$.

distributions, blacklisting is effective only if implemented quickly, i.e. before websites have had sufficient time to recover. The likelihood of remaining infected for t time steps is $(1-\gamma)^t$, which becomes exponentially small for large t. For example, once $\beta > 40$, the steady state expected exposure is very close to the theoretical value with no interventions (around 0.091). Thus, for larger β, most infections will resolve before infected websites are blacklisted. The precise relationship between γ and β is given by Equation 5.7.

The results of varying the depreferencing parameter, σ, are shown in Figure 5.11

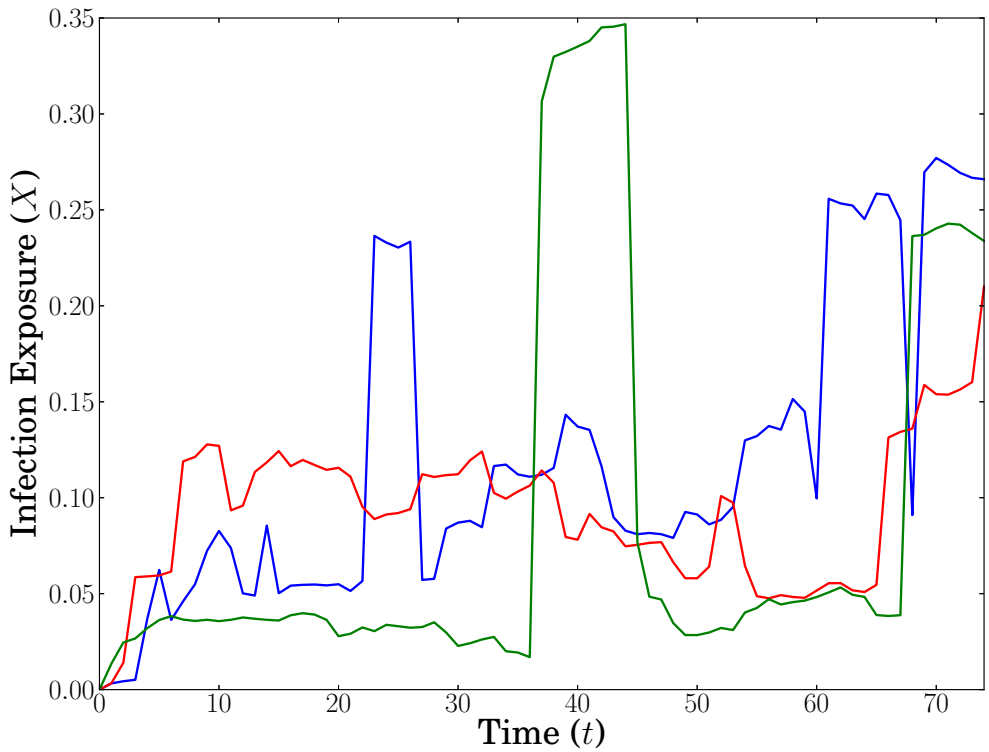

Figure 5.7: Variation in infection exposure in individual runs for power-law distributions. Here the parameters are held constant to illustrate the variation between any two simulation runs. $\rho = 0.01$, $\gamma = 0.1$

and Figure 5.12. Because proportional depreferencing of popularity has an exponential impact on the ranking (Equation 5.2), even large values of σ can reduce infection rates significantly, for example, when $\sigma = 0.9$, the steady state client infection rate is half of the baseline value.

Depreferencing gives finer control to search engines, because adjusting σ should be relatively easy, unlike trying to reduce β, the control parameter for blacklisting. This finer control might allow for algorithms that produce more false positives (which in turn would reduce the number of missed infections), because the effects of being

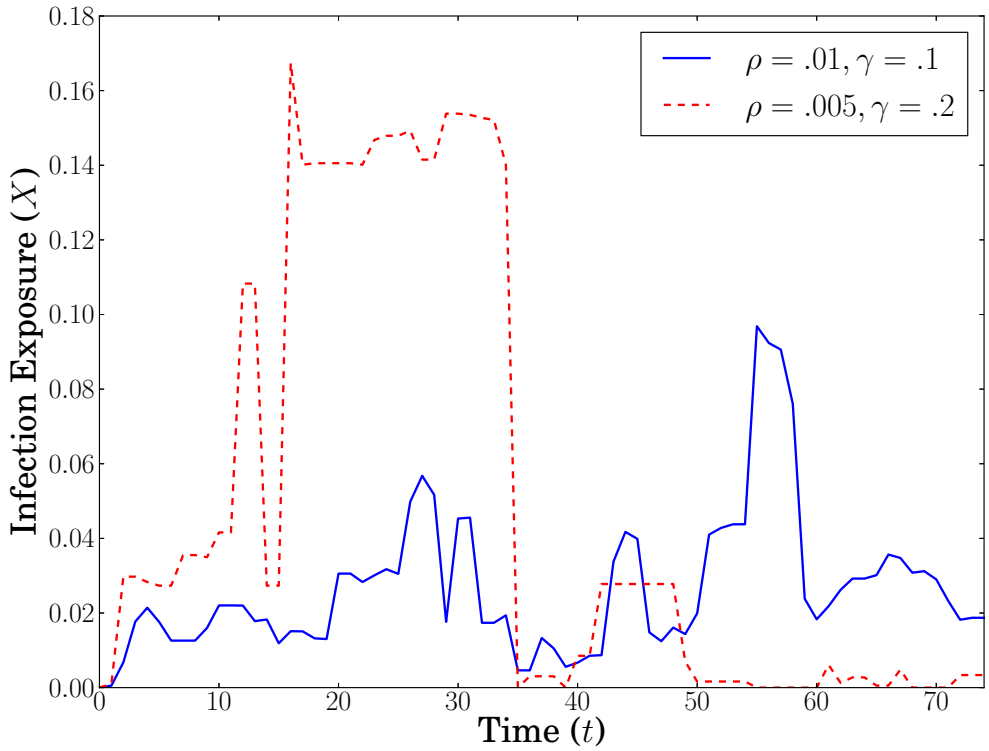

Figure 5.8: Variation of infection exposure in individual runs for power-law distribution of website popularity. Simulation parameters differ here to show how even increased recover and decreased infection can appear to have worse outcomes

mislabeled as infected could have far less impact on a website that was moved down in the search rankings rather than being blacklisted.

5.5.3 False Positives

Depreferencing makes it feasible to use imprecise detection algorithms that trade faster detection for higher false positives. In our model, this would translate into a higher value for f, the false positive probability. Figure 5.13 and Figure 5.14 explores the impact of f on the change in traffic loss due to false positives. Once again, a large

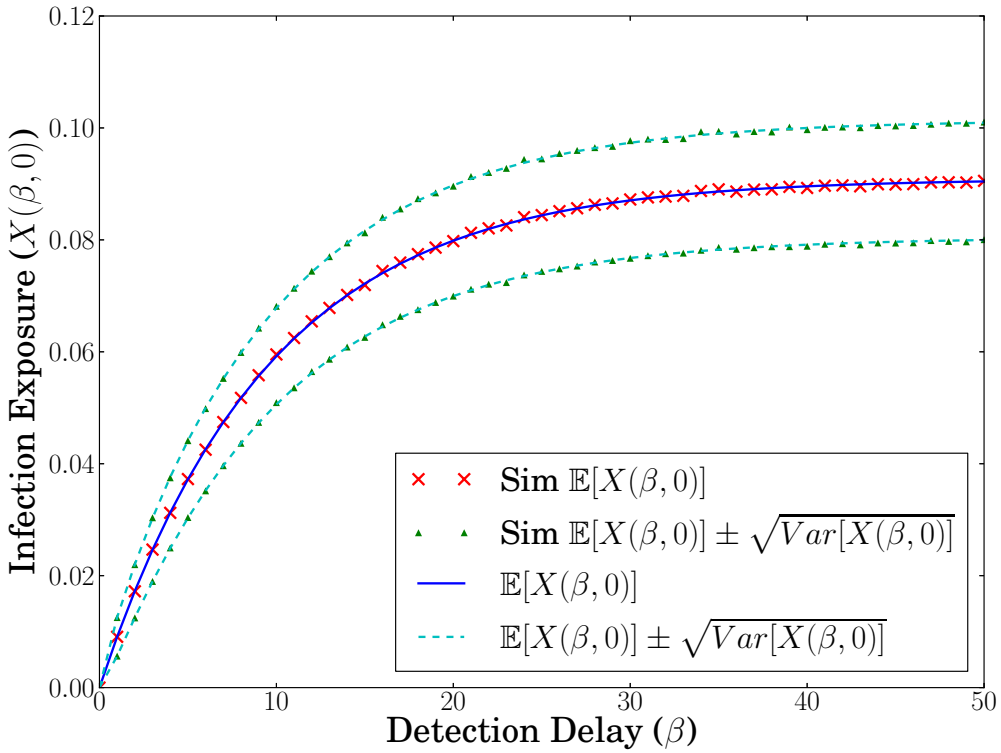

Figure 5.9: Steady state client exposure to infection with uniform traffic for various detection delays, β, with $\sigma = 0$.

variance in the website popularity distribution has a large impact on the outcome, i.e. the traffic loss. Further, as can be seen in Figure 5.13 and Figure 5.14, reducing the false positive rate is only worthwhile if it can be dropped below a certain value (in this particular example, around 0.2); when f is high enough, every website is mainly in the infected or falsely infected state, and rarely in the uninfected state.]

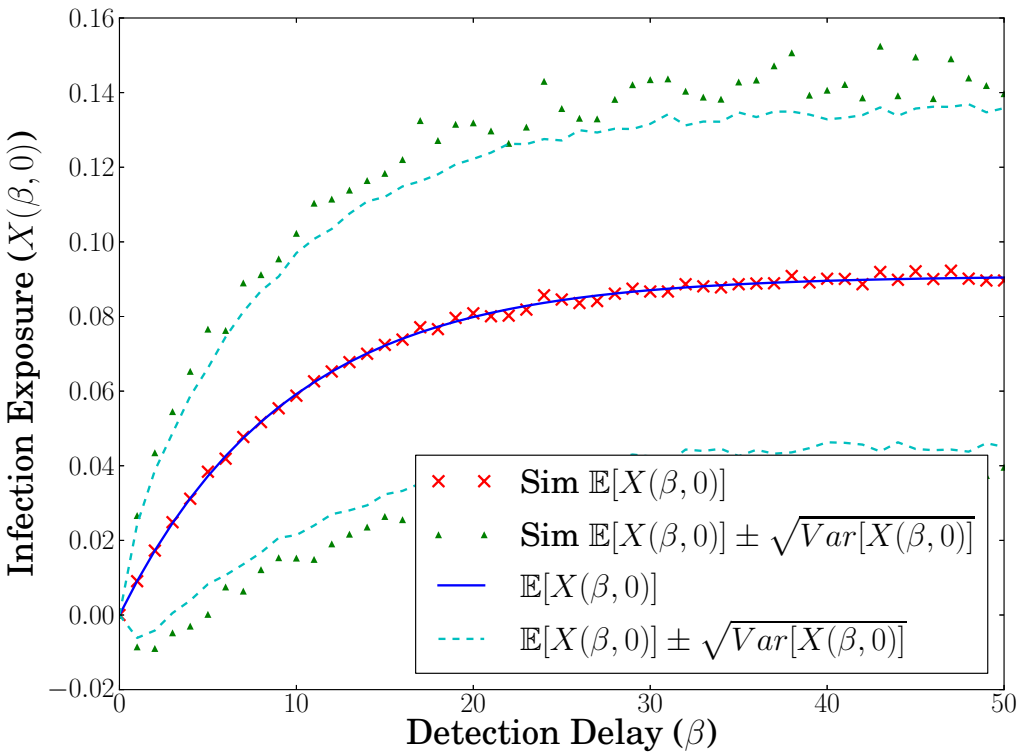

Figure 5.10: Steady state client exposure to infection with powerlaw traffic for various detection delays, β, with $\sigma = 0$.

5.5.4 Exploring the Parameter Space

Figure 5.15 and Figure 5.16 show how the expected infection exposure and traffic loss change as the parameters σ and β vary from a base setting of $\beta = 10$ and $\sigma = 0.5$. We can see from the solid line at the critical value in Figure 5.15 that changing the depreferencing parameter, σ, can only correct for a small increase in β, up to $\beta = 11$. Beyond that, the expected exposure increases, regardless of the setting of σ. The value of σ only starts to have a large positive impact if the detection delay, β, drops significantly. We see similar results for the change in expected traffic loss, as shown in Figure 5.16. Once again, only the smallest increases in β can be compensated

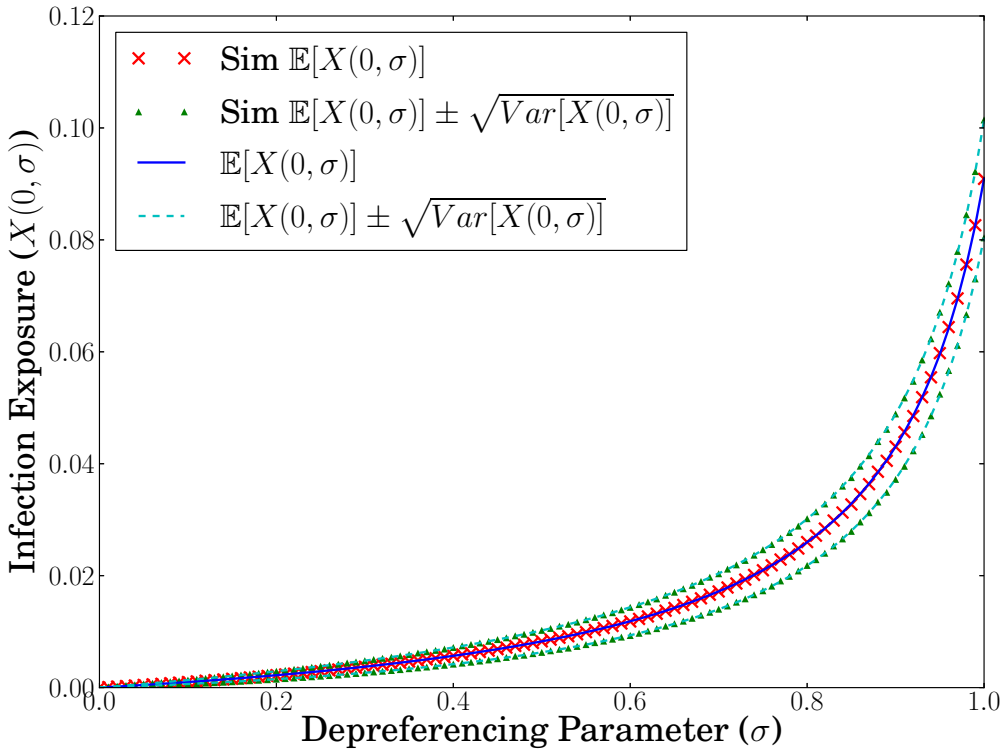

Figure 5.11: Steady state client exposure to infection with uniform traffic for various depreferencing adjustment values, σ, with $\beta = 0$.

for by increasing σ. However, lack of compensation means a decrease in traffic loss, which is a desirable outcome. We also see that it is easy to adjust σ to ensure that the traffic loss does not increase for almost every change in β.

It is clear that a faster response (reducing β) will reduce the infection exposure rate, and any potential traffic loss can easily be compensated for by changing σ. However, a faster response may be less accurate and result in a higher false positive rate, f. We explore this idea by again calculating the infection exposure with base values $\beta = 10$ and $\sigma = 0.5$, and then calculating the critical value σ_X needed to maintain the same infection exposure rate for a variety of β' values. We then measure

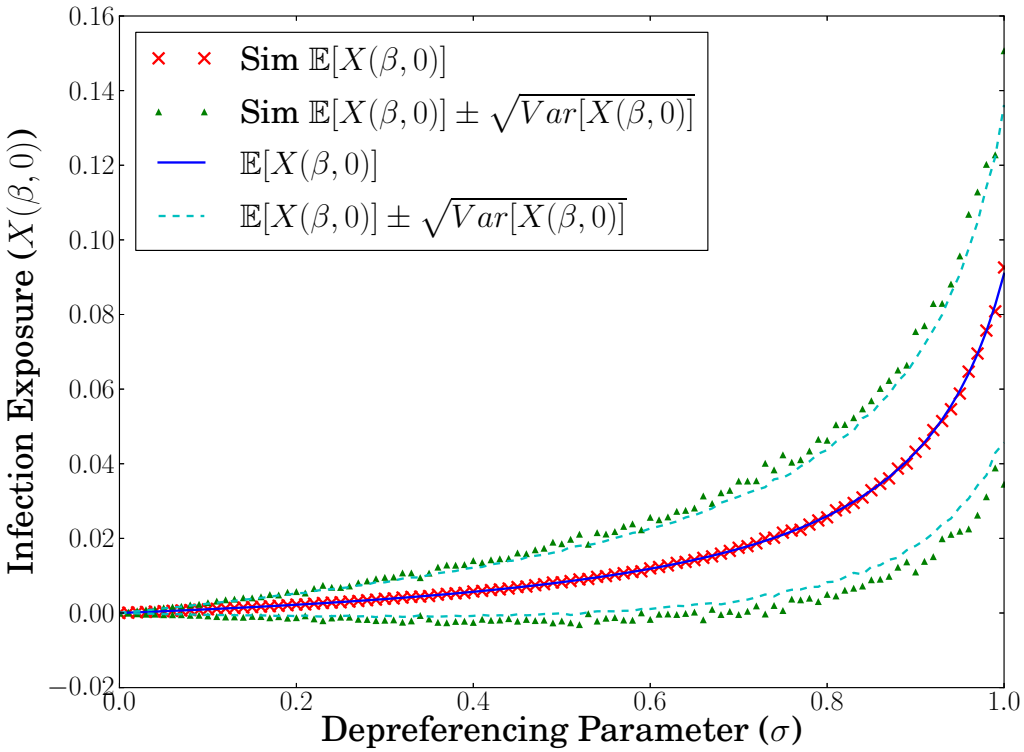

Figure 5.12: Steady state client exposure to infection with power-law traffic for various depreferencing adjustment values, σ, with $\beta = 0$.

the change in traffic loss $\mathbb{E}[L(10, 0.5)] - \mathbb{E}[L(\beta', \sigma_X)]$ for a variety of false positive rates. The results can be seen in Figure 5.17. Generally, a decrease in detection delay, β, increases the traffic loss for a constant false positive rate. If the false positive rate also goes up as β decreases, the problem is even worse. However, if the false positive rate can be kept sufficiently small (below 0.1 in this example), then there is flexibility to decrease the delay without a major increase in traffic loss.

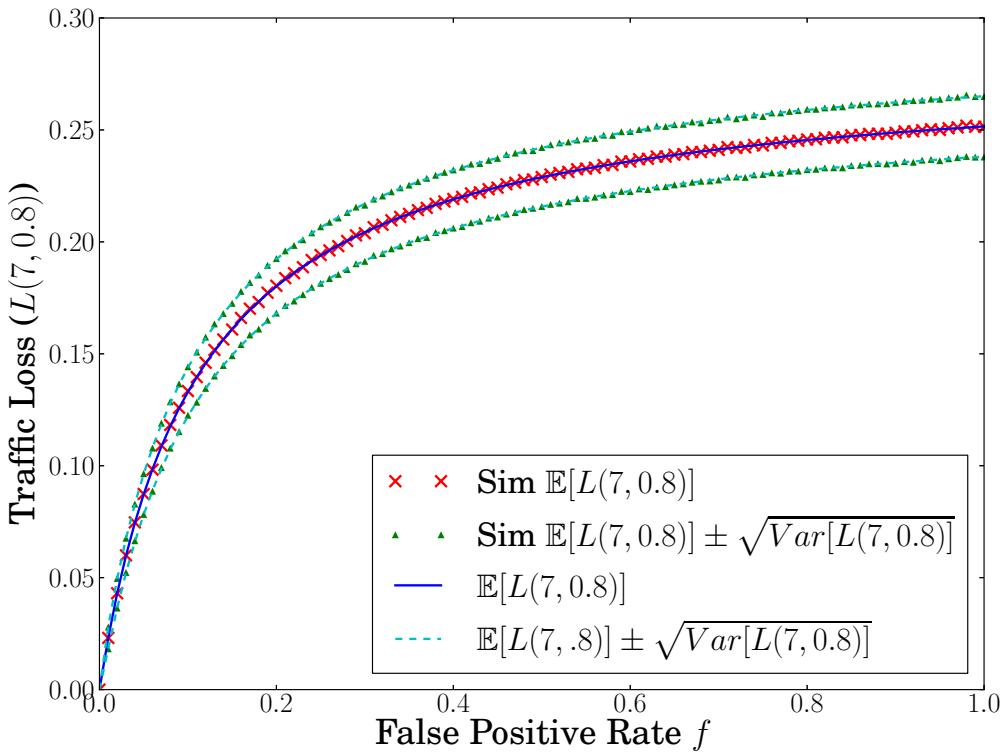

Figure 5.13: Steady state normalized traffic loss for various false positive rates. Each data point is the average of 1000 runs, with depreferencing parameter $\sigma = 0.8$.

5.5.5 Parameter Distributions

We also explored the effect of drawing the parameters ρ, γ, and f from distributions, instead of using constant values, but did not find an easily obtained analytic form for the distribution or moments of X or L. Preliminary simulations, however, suggest that the distribution of these values follows the joint distribution of Equation 5.7 and Equation 5.9 (results not shown). Further, our earlier result that there is large variance in the measured outcomes is observed when ρ and γ are related to ω_i. Further investigation into interactions among these paramters is left for future work.

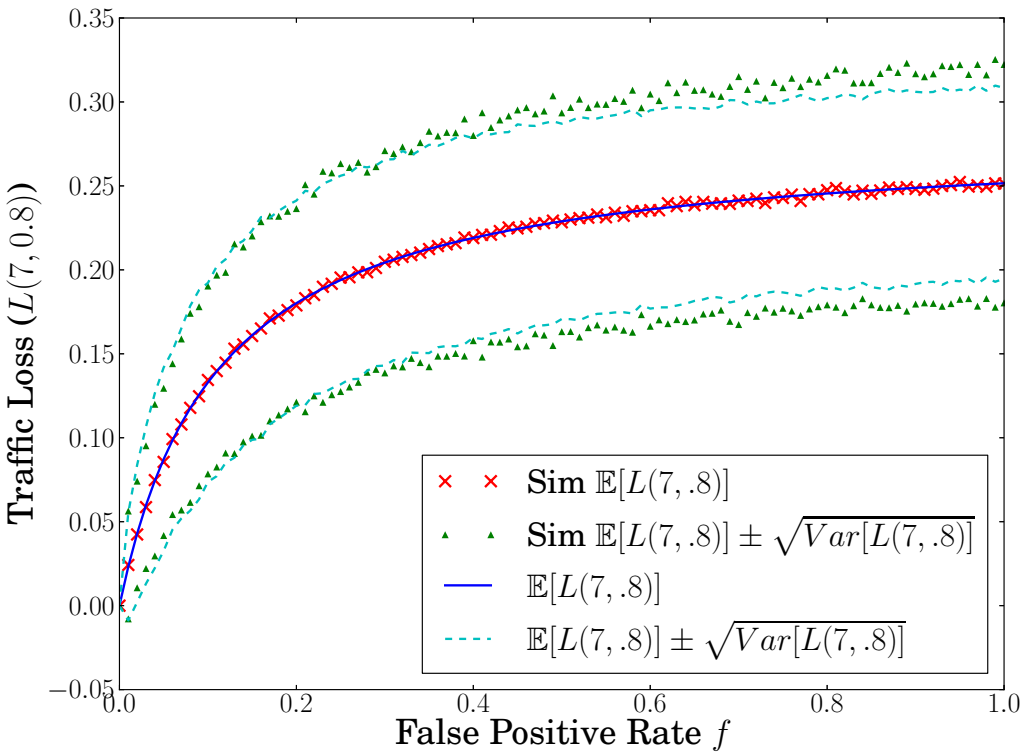

Figure 5.14: Steady state normalized traffic loss for various false positive rates. Each data point is the average of 1000 runs, with depreferencing parameter $\sigma = 0.8$.

5.6 Related Work

There are many approaches to combating web-based malware, including the use of virtual machines or kernel extensions to check for suspicious changes to the operating system [195, 287, 221, 179], emulating browsers to detect malicious JavaScript [58, 64], and detecting campaigns that promote compromised sites to the top of search results [131]. No technique is completely effective at disrupting web-based malware, according to a study of Google's data over more than four years [224]. In our view, one limiting factor is the choice of conservative approaches that minimize false positives at the expense of speedy detection. For example, Provos et al. [221] choose to

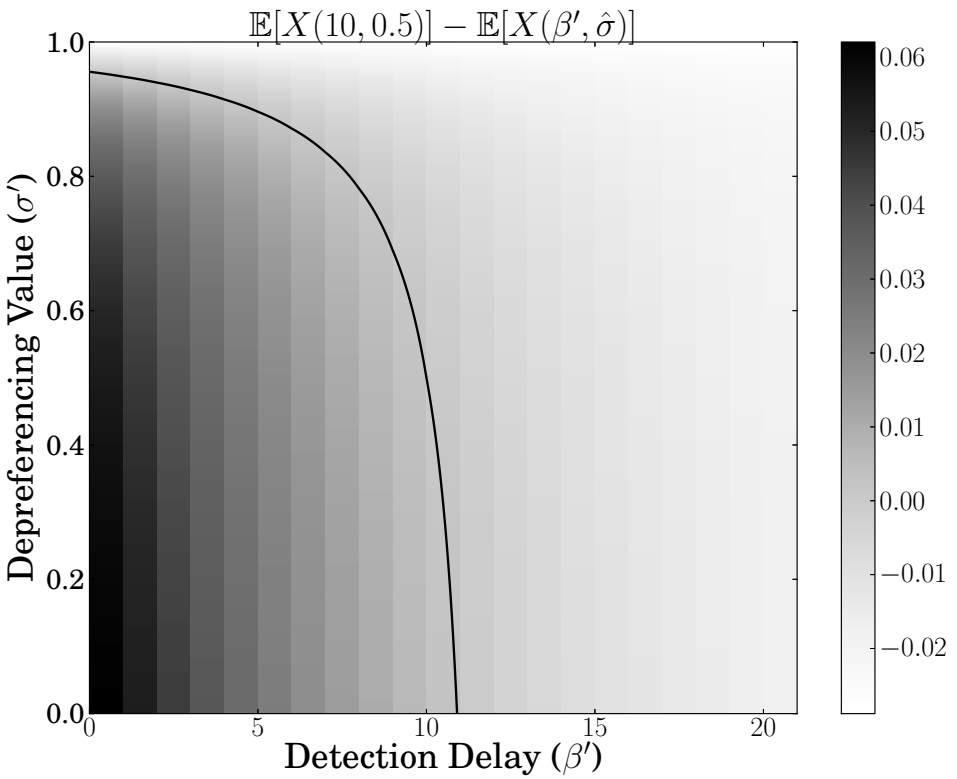

Figure 5.15: Change in expected infection exposure when parameters β' and σ' vary from a base value of $\beta = 10$ and $\sigma = 0.5$. The solid line corresponds to the critical value $\sigma' = \sigma_X$.

minimize false positives in a system that allows explicit trade-offs between false and true positives.

Depreferencing of search results is an example of a graduated response, which is different from the binary, all-or-nothing, response methods, such as blacklisting, that are usually taken in cybersecurity. An early implementation of graduated response was a Linux kernel extension called pH [253], which responded to anomalous system call patterns by delaying subsequent system calls in the offending process. Other graduated responses operate by slowing down, or throttling, outgoing requests [294,

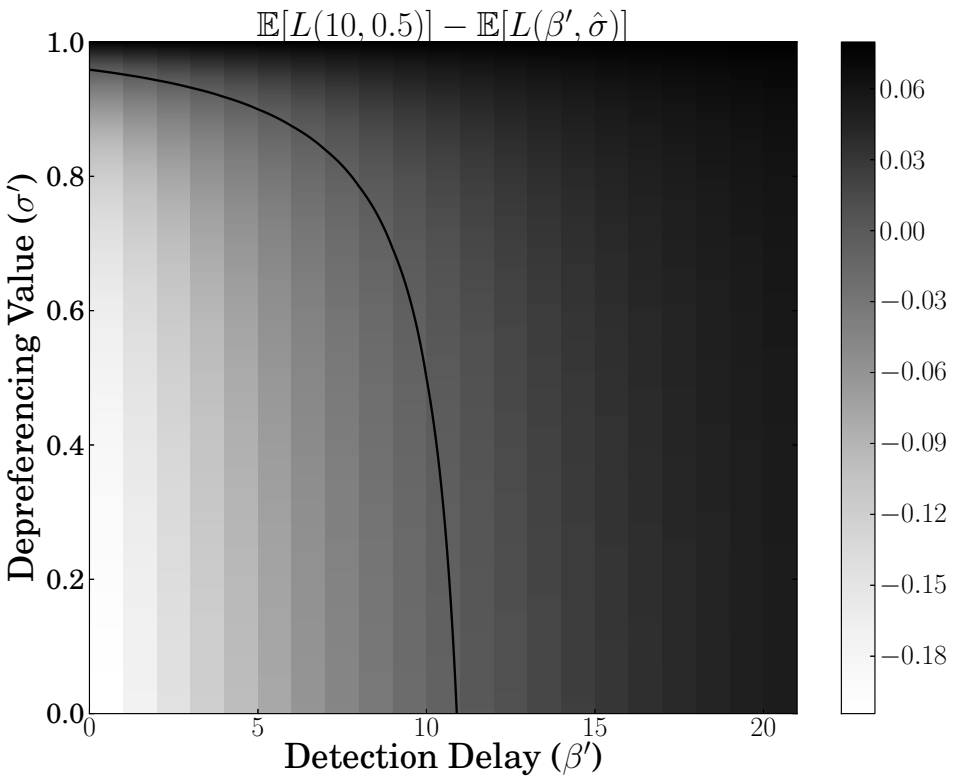

Figure 5.16: Change in the expected traffic loss when parameters β' and σ' vary from a base of $\beta = 10$ and $\sigma = 0.5$. The solid lines correspond to the critical values $\sigma' = \sigma_L$.

123] in active networks [115], Domain Name Service [297], Border Gateway Protocol [138], and peer-to-peer networks [98]. However, this is the first work we are aware of that uses a graduated response outside of the time domain.

Several studies have focused on alternative intervention strategies, which could potentially be generalized using our depreferencing method. For example, Hofmeyr et al. modeled responses available to ISPs [117]. Other researchers have identified suitable intervention strategies based on empirical research, which might also be amenable to depreferencing. For example, Levchenko et al. [167] found that

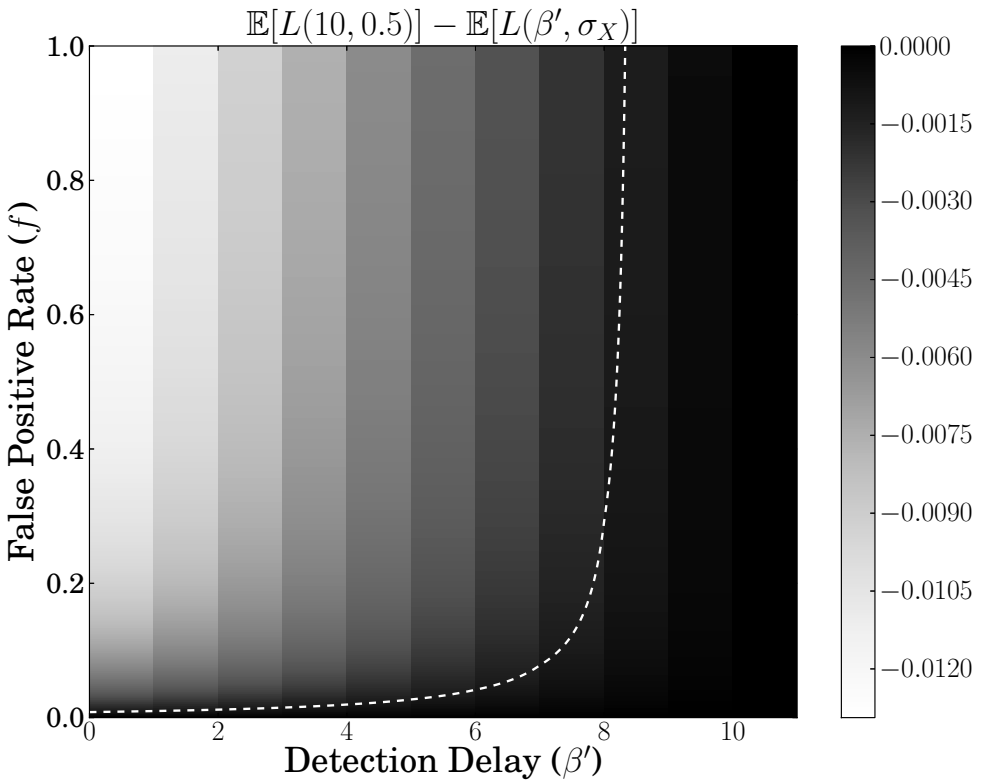

Figure 5.17: Change in expected traffic loss when expected infection exposure is kept constant, i.e. $\sigma' = \sigma_X$, the critical value. The base for comparison is $\beta = 10$ and $\sigma = 0.5$. The dotted line corresponds to a value of -0.1, i.e. an increase in traffic loss of 10%.

criminals relied on just three payment processors to collect money from victims, which led the authors to recommend targeting the payment processors as a low-cost intervention. Similarly, Liu et al. [173] empirically measured the effectiveness of pressuring registrars to suspend spam-advertising domain names. In a related intervention, Google has successfully pushed ad-filled sites down the results by changes to its search-ranking algorithm [193], suggesting that a similar effort to depreference malware-infected sites is technically feasible.

5.7 Discussion

A general theme of this research is the emphasis on modeling. Modeling is a cost-effective way to explore intervention strategies, including investigating novel ideas, without the expense of first implementing them. As our results show, modeling can be particularly helpful for understanding long-term trends in processes with high variance, where direct experimentation can be misleading. Thoroughly testing the interventions we explore in this chapter would likely require an unreasonable amount of time and money for any search provider.

To the best of our knowledge, the idea of website depreferencing to prevent the spread of malware has not appeared in the literature before. Although we believe that depreferencing is technically feasible[8], other issues may arise with this type of response. For example, a policy that explicitly tolerates false positives could trigger accusations of bias against search engines.[9] In the future government regulations may require search providers to enact measures which both avoid bias and protect users from malicious software.

Another issue is how depreferencing might be gamed. For example, there could be an incentive to deliberately infect competitors' websites, or cause them to appear infected, so their search rankings are demoted. Such industrial sabotage may in fact already happen. However, the scope for it could increase if less precise, false-positive tolerant detection mechanisms are used.

Depreferencing may have other advantages over traditional blacklisting. As web content and attacks become more and more sophisticated, it will may become more

[8]For example, the Google Penguin update uses a form of depreferencing to decrease rankings for websites that violate Google's quality guidelines (http://insidesearch.blogspot.co.uk/2012/04/another-step-to-reward-high-quality.html).

[9]The European Union is already investigating accusations that Google abused its power by preferring its own results over rivals. See `http://www.time.com/time/business/article/0,8599,2034138,00.html`.

difficult to distinguish infection from non-infected states. For example, websites may host malicious advertisements in a frame, making only a part of a website infected. Moreover, sophisticated attacks may hide themselves periodically. In these cases infection is not a binary state, but could be measured by degree. The degree of infection could then be incorporated into the depreferencing parameter allowing for a proportionate response. If the computational cost of detection becomes a bottleneck, fast, less precise methods will become necessary.

We have made several simplifying assumptions that we believe are reasonable in the absence of more detailed information. For example, we assume that website infection and client infection probabilities are independent. In reality, this may not be the case. One variety of drive-by-download malware steals the login credentials of users who administer websites, enabling the malware to spread to those websites. Hence, when a client is infected, the probability of infecting one or more websites increases, corresponding to a change in ρ. We have chosen not to model this form of malware spread because it has been observed only in a handful of outbreaks (e.g., one Zeus variant in 2009 [82]).

Another assumption is that the distribution of website popularity is time invariant, which is true in general, although the popularity of individual websites can vary over time [147]. However, the popularity of infected websites may change over time when attackers attempt to promote compromised websites in search-engine rankings [131]. In future, if sufficient information can be attained, it may be possible to accurately model this aspect. We believe, however, that even with more accurate information, the heavy tailed nature of popularity will cause similar heavy tailed behavior in infection exposure and traffic loss.

We also assume that users treat all search results as equal, differing only by ranking. This is likely untrue in the case where users are searching for a specific website, and there could be other effects, such as an abrupt cut-off after the first

page of results. More data is needed on the exact nature of user-responses. It is possible that depreferencing could be implemented not a as a reduction in ranking, but by some other mechanism. For example, a search engine could provide multiple warnings with different degrees of difficulty to navigate for infected websites. We leave these aspects to future work.

We make the assumption is that the recovery rate from false positives and actual infections is the same. It is plausible that actual infections could exhibit other signs which would warn website administrators of an infection, speeding recovery. Conversely, malware could actively hide itself, slowing recovery. Moreover, recovery rates might changed based on the type of intervention taken. In blacklisting, a sudden loss of traffic might warn administrators about infections faster depreferencing's slow reduction in traffic. This might require a warning from the search provider to a website administrator that a websites traffic is being altered.[10]. This is especially important in the case of if higher false positive rates are to be tolerated. Different recovery rates based on these scenarios could be incorporated into the model, but we leave this analysis for future work.

Another area of future work would be to focus on infections that spread in a general network environment where a referral service (such as search) plays a key role. Similar interventions could be applied when infections are spread from website to website, rather than simply exposing a client population. This could be a particularly good model for controlling infections of malicious software in online social networks.

In our analysis and modeling we disregard the effect of false negatives, primarily because we assume that the response methods we explore use the same detection mechanisms, subject to the same false negative rates. Usually, in real detection systems, reducing the accuracy of the system by increasing false positives usually leads to a *decrease* in false negatives, a feature which gives rise to the traditional

[10]Administrators are likely to ignore automated warnings[278]

ROC curve. We have insufficient data to model this effect, but it suggests that the depreferencing mechanism could have additional benefits beyond those shown by the model: increasing tolerance of false positives could also improve the rate of detection of compromised sites.

Our focus in this research has been to develop a plausible model that allows us to assess the impact of different interventions on the spread of drive-by-download malware. Our goal is to show that modeling can be a useful tool for search providers to use when considering different interventions. We do not have access to data that could enable us to make quantitative predictions about interventions. We expect search providers to have much more relevant data, especially information on the distribution of website popularity, the efficacy of infected website detection, the recovery times for infection and the behavior of users.

5.8 Summary

We proposed and explored a novel intervention strategy, called *depreferencing*, where a possibly infected website is moved down in the search results, rather than outright blacklisted. Depreferencing may be an attractive alternative to blacklisting for search providers because it allows them to use less precise detection methods with higher false positive rates, potentially increasing the speed of response to infection and reducing the cost of detection. These results imply great difficulty in determining empirically whether certain website interventions are effective, and it suggests that theoretical models such as the one described in this chapter have an important role to play in improving web security.

Chapter 6

Strategic Aspects of Cyber-Attribution[1]

6.1 Introduction

International conflict and espionage are no longer confined to physical space. The Internet provides an avenue for nation states to steal industrial secrets [112], gather important intelligence [200], alter the results of elections [272], and even cripple physical infrastructure [85]. However, the evidence of these actions is more ephemeral than those in the physical world. Digital records can easily be copied, altered, or deleted; identities can be faked; and attacks can be made to appear to be the result of accidents or incompetence. Furthermore, the ability to coordinate such attacks is not confined to nation states. Any sufficiently technically adept group, be it patriotic

[1]Portions of the material in this chapter were developed in collaboration with Robert Axelrod and Alexander Furnas. Robert Axelrod conceived the original Responsibility Game and provided many helpful pointers to the literature. Alexander Furnas aided in the analysis of the Responsibility game. I conceived and analyzed the Asymmetric Prisoner's dilemma and the Attribution Game, and I prepared the final written presentation.

hackers, terrorists, or organized criminals can cause damage on the same scale as many nations.

How should nation states respond to new cyber events given this context of comparatively difficult attribution? The United States has indicated that the forensic problem of determining the technical origin of an attack has been largely solved [47]. But if this is true it raises the obvious question: Why hasn't the US had a stronger response to cyber incidents such as theft of intellectual property [300] and the leak of personal information of 21.5 million federal employees [67]?

As in the previous chapter, data on cyber-conflict are generally difficult or impossible to obtain, as national security concerns usually prevent the sharing of information openly. In situations such as these abstract modeling becomes a critical tool in developing a quantitative approach to these problems. In this chapter, we present three game-theoretic models, which incorporate unique aspects of cyber conflict and lead to surprising implications for strategies in this domain.

Specifically, we incorporate three important aspects of cyber-conflict into our game-theoretic models. First, the relationship between the perpetrators of an attack and the nation where they reside may not be clear to the victim. Therefore it can be challenging for a victim country to hold another country responsible for an attack they can reasonably deny orchestrating. For example, the large Distributed Denial of Service (DDoS) against Estonia's Internet infrastructure in 2007 was originally blamed on Russia who denied responsibility[2] [241]. We explore this aspect in the model in section 6.2.

Second, victims of attacks may not have the option of a proportional response, i.e. a retaliatory attack in the same domain with similar consequences to the initial attack. For example, when North Korean attackers compromised Sony Pictures En-

[2]A Estonian citizen of Russian ethnicity was held responsible for the attack and fined

tertainment and exfiltrated and leaked confidential emails and intellectual property, the U.S. had no comparable target within North Korea [240]. The US's only option was to either accept the attacks or proceed to retaliate with a potentially disproportionate response leading to further escalation. In this case the U.S. chose to pursue economics sanctions [239]. The game outlined in section 6.3 demonstrates that even in two player games, asymmetric capabilities can lead one player rationally prefer to tolerate small attacks rather than escalate the conflict.

Even when symmetric responses are available, asymmetric uncertainties can cause similar instability. In the last game, we find that increasing an adversary's ability to attribute attacks may decrease the likelihood of escalating conflict. Moreover, highly asymmetric attribution ability to attribute attacks may lead to situations where the player with the higher attribution ability is forced to continually respond to attacks from adversaries with punishments, leading to lower than optimal payoffs. We explore this in the final game presented in section 6.4.

The three game-theoretic models presented in this chapter incorporate each of these different aspects to varying degrees. The first model incorporates all three, but is a challenge to analyze. We show that while the game can be analyzed, the full results of the analysis only hint at the strategic lessons for cyberconflict. However, small aspects of the results give clues to details we then explore in two simplified models. These three models do not describe the complete cyber conflict landscape. However, they do illustrate several important aspects of cyber attack which has yet to be explored.

6.2 The Responsibility Game

In this section we describe the Responsibility Game (RG), a game theoretic model of the types of conflicts discussed in section 6.1. This game has players who may not

Table 6.1: Prisoner's Dilemma Payoffs. We denote Cooperate C and defect D. The payoffs are denoted X, Y, with X being A's payoff and Y being B's payoff.

		B's action	
		Cooperate	Defect
A's action	Cooperate	0,0	$-L$,G
	Defect	G, $-L$	$G-L$, $G-L$

be able to directly interact with other players, players who have asymmetric attacks, and uncertainty about the relationship between players.

6.2.1 Players, Actions, Payoffs, and Information

The RG has three players, which we denote A, B and C. As noted in section 6.1, we are investigating conflicts which involve two main parties who are peers, such as the US and China, denoted A and B respectively. The game also includes a third player, C, who is affiliated B, or shares similar goals as B. This group could be patriot hackers, or simply criminal hackers. The game is represented in figure 6.1.

The RG consists of three separate components: a Prisoner's Dilemma (PD) component played between A and B, an attack component played by A and C, and a penalty component played by B and C. We parameterize the PD played between A and B slightly differently than it is traditionally presented [12]. In our version, players receive gain G from defecting and lose L from being defected against. Neither player receives explicit payoff from mutual cooperation. The payoffs for the PD can be found in Table 6.1. This is equivalent to the traditional formulation when $T = G$, $R = 0$, $P = G - L$, and $S = -L$. Here, the assumption that $T > R > P > S$ is separated into the two requirements that $G > 0$, $L > 0$, and $L > G$. We use this formulation because it maps more naturally to the cyber domain where players are attempting to steal, for example, industrial secrets from one another [300].

The second component of the game is played between A and C. In this component

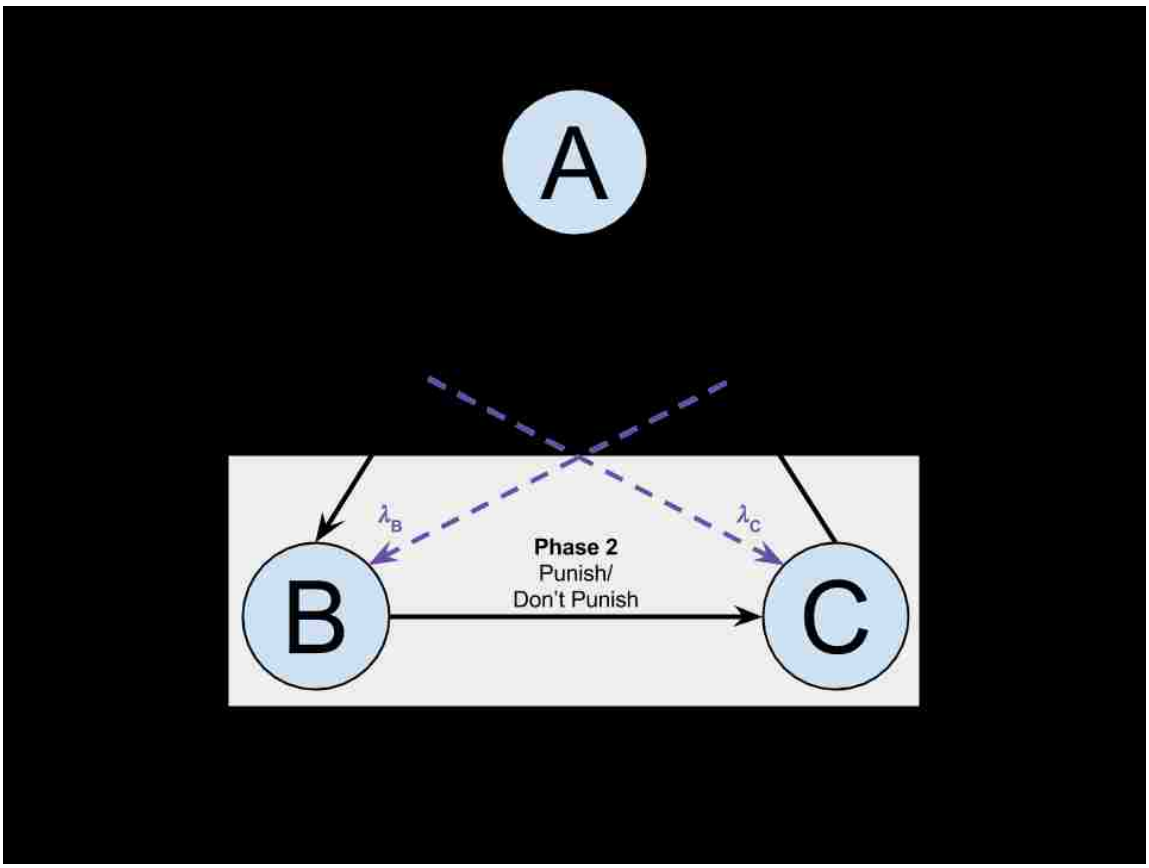

Figure 6.1: A diagram of the Responsibility Game (RG). B and C are assumed to have mutual interests and partially share their payoffs through the values λ_B and λ_C. Actions are visible to all players, except that A cannot see the interactions between B and C.

A is passive and takes no action against C. C can, however, choose either to attack(A) or not attack(NA). When C attacks, A takes damage $-d_{CA} \leq 0$, and C gains utility $\tau \geq 0$. This component corresponds to situations in which C is able to attack A, but A is unable to respond as is the case with the United States and possibly state-sponsored hackers operating out of China.

The final component of the game is played by B and C. In this component B can penalize C, e.g. if C is a cybercriminal violating B's laws. Specifically, B can choose

to penalize(P) or not penalize(NP) C. When B penalizes C, it costs B $-k \leq 0$, and does damage $-d_{BC} \leq 0$ to C.

We assume that all players know the actions taken by all of the other players in each component with one exception: A cannot observe whether B punishes C or not. This models cases where B may have partial control over C, but chooses not to broadcast the nature of its relationship with C.

Shared Payoffs

In addition to the payoffs obtained by players in each component of the game, we add a feature that represents the relationship between B and C. Because B and C share similar goals, we assume that each benefits from the action of the other in their interactions with A. We assume that B benefits from attacks by C by receiving some of C's payoff, represented by λ_B, with $0 \leq \lambda_B \leq 1$. Specifically, we assume that B receives $\lambda_B \tau$ when C attacks, and no payoff when C does not attack.[3] Similarly, we assume that C's payoff from the component between B and C, represented by the quantity $0 \leq \lambda_C \leq 1$. We assume that C prefers that A do worse in the PD, and receives $\lambda_C - (PD_A)$, where PD_A is A's payoff in the PD. This means that C receives no payoff for mutual cooperation, $\lambda_C L$ when B defects and A cooperates, $\lambda_C(-(G-L))$ for mutual defection, and $\lambda_C(-G)$ when A defects and B cooperates.[4] λ_B and λ_C represent the strength of the relationship between B and C, and not a fractional shared payoff.

[3] λ_B is not a fraction, that is, when B receives $\lambda_B \tau$, *C does not* receive $(1-\lambda_B)\tau$, but continues to simply receive τ.

[4] This shared payoff could be represented a number of different ways. C simply prefer that B defect against A, or could prefer the advantage of B over A. We leave these alternative models for future work.

Component Order

The RG is infinitely repeated with the payoffs in each round discounted by $0 \leq \delta \leq 1$. A round consists of a single play of each of the three components. We assume that the RG's components are played in the following order: C chooses whether to attack A or not; A and B play a one round PD, and simultaneously B chooses whether to punish C or not.

6.2.2 Actions and Equilibrium

We do not give a complete analysis of the game here (see Appendix A and B for details), because depending on the parameter values nearly all possible strategies can be a Nash equilibrium. But, there are a few important features of this game. First we consider in equilibria in which all players maintain *peaceful* play, that is A and B cooperate, C does not attack A, and B does not punish C. Next, we consider situations when C attacks A, but A has no incentive to retaliate.

In the infinitely repeated version of the game, one of the conditions under which peace is maintained is given by the equation:

$$\delta \geq \frac{G}{L - \lambda_B \tau} \tag{6.1}$$

While this is only one condition for achieving peace, it is an illustrative one. If this condition is met, then B has no incentive to pre-emptively defect against A. We note that as λ_B grows the incentive for cooperation decreases.

Now, suppose a situation in which A is uncertain about the relationship between B and C, and, in particular, the parameter λ_B. If A underestimates this parameter, she will anticipate future cooperation from B and will cooperate herself, even as B

prepares to defect. Conversely, if A overestimates λ_B it may pre-emptively defect against B as A's analysis of the game will lead it to believe that B will defect. In the latter scenario B would actually prefer A's estimate of λ_B to be more accurate to avoid defection.

This illustrates an important aspect of cyber attacks. The victim of an attack may not know the exact relation between the perpetrators and the country in which they reside. This was the case in the 2007 attacks on Estonia's infrastructure [241], and the Office of Personal Management breach [199]. Underestimating this relationship could cause the US to fail to punish its attackers, while overestimation could lead to unnecessary escalation of cyber attacks. In the next two sections we explore simplifications of the RG.

6.3 Asymmetric Prisoner's Dilemma

This section presents a simplification of the RG. This simplified two player game illustrates an important outcome: in some situations it is rational for the victim of an attack to cooperate in the face of minor attacks, when the do not have a proportional response. This may be the case in many cyber conflict scenarios.

6.3.1 Asymmetric Prisoner's Dilemma

Here we describe a simplified version of the responsibility game, which is more amenable to analysis. The Asymmetric Prisoner's Dilemma (APD) involves two players (A and B) playing a PD with an added asymmetry: player B has a choice of two defection levels. Specifically, A and B can *cooperate (C)* or *defect (D)* as in the standard PD, and player B has the third option of *small defect (d)*. The payoffs in this game differ slightly from those in the PD presented in section 6.2 (See table 6.1).

In the APD, we set $G = 1$, which is equivalent to scaling all the payoffs in table 6.1 by G. As before mutual cooperation earns both players 0, a defection earns the defecting player $G = 1$ and the victim $-L$ with $L > 0$. Mutual defection earns both players $G - L$. As before we assume that $L > G = 1$ to enforce the classic PD conditions. If B chooses to play small defect, she gains $g < G$ and A experiences loss $-\lambda L$ with $0 < \lambda < L$. The game can be characterized by four parameters, all between 0 and 1, namely: $\frac{1}{L}$, δ, g and λ. This simplifies our the parameter space. The payoff table for all combinations of actions can be seen in Table 6.2.

Table 6.2: Payoffs for the asymmetric PD. Each cell is X,Y where X is A's payoff for the given action set, and Y is B's payoff for the given action set

		B's Actions		
		Cooperate(C)	small Defect(d)	large Defect(D)
A's Actions	Cooperate	0,0	$-\lambda L$,g	$-L$,1
	Defect	1,$-L$	$1-\lambda L$, $g-L$	$1-L$, $1-L$

The game is played repeatedly, with each players payoff each round discounted by a factor $0 < \delta < 1$. That is for each round, i, players choose actions, and receive payoff $\delta^i X$, where X is the payoff of the chosen actions in table 6.2. Each player moves simultaneously each round, and can see all moves made in the game, and can remember all previous moves of both players.

6.3.2 Strategies and Equilibrium

As in the RG, we consider pure strategies of the following form: players cooperate, until the other player defects. After the defection, the victim will defect for n rounds to 'punish' the first defection. These types of punishment phase strategies are often studied as equilibrium strategies in the iterated PD, e.g. [80].

We study this class of strategies because we are interested in situations where A can hold B responsible both for a small defect and a large defect. That is, when does

there exist an n with which A can threaten B, such that B will be deterred from small defecting.

Two conditions need to be true for such an n to exist. The threat must be *effective* and *credible*. A threat is effective if B's payoff is lower under the punishment than it would be if B had cooperated in all rounds. A threat is credible if A's payoff is higher carrying out the punishment than it would be if A simply allowed B to defect (small or large) in all rounds.

We assume that during A's n rounds of defection, B also plays D so as to minimize its losses, and that after the punishment phase both players return to C. However, in real world situations we could imagine that while A plays D, outside pressure forces B to play C throughout the punishment phase. It could be that B plays d during the punishment phase, i.e. denying responsibility for the small defects. We explore this in section 6.4.

6.3.3 Punishment is Effective

From A's perspective we need to consider whether punishments against B are effective in the case of small and large defection. It might also be in A's best interest to defect on the first round. Specifically, if δ is small then it is possible that the best strategy for both A and B is to simply defect on the first move. Specifically, A or B can threaten n rounds of defection following a large defect from either player. Since the situation is symmetric, we need only consider one player's perspective.

Theorem 6.3.1. *Punishment is effective against large defections if* $\delta > \frac{1}{L}$.

Proof. Cooperation has a greater payoff than defection if:

$$
\begin{aligned}
0 &\geq 1 + (1-L)\sum_{i=1}^{n} \delta^i \\
0 &\geq 1 + \frac{(1-L)\delta(1-\delta^n)}{1-\delta} \\
(1-\delta^n) &\geq \frac{1-\delta}{(L-1)\delta}
\end{aligned} \tag{6.2}
$$

When equation 6.2 holds, the threat of n rounds of punishment are an effective deterrent against large defection. If equation 6.2 does not hold, i.e. there is no n such that satisfies the condition. This simplifies to $\delta \leq \frac{1}{L}$[5]. □

We next examine whether effective punishments exist for small defection. That is can a punishment phase from A dissuade B from small defecting in a single round.

Theorem 6.3.2. *A has an effective punishment against small defections if* $\delta > \frac{g}{L-1+g}$.

Proof. We can express the conditions under which the punishment is effective against small defection through the following inequality:

$$
\begin{aligned}
0 &\geq g + (1-L)\sum_{i=1}^{n} \delta^i \\
0 &\geq g + \frac{(1-L)\delta(1-\delta^n)}{1-\delta}
\end{aligned} \tag{6.3}
$$

This simplifies to two equivalent statements:

[5]We note that this is equivalent to condition for mutual cooperation in the iterated PD using our transformation.

$$\frac{g(1-\delta)}{(L-1)\delta} \leq (1-\delta^n) \tag{6.4}$$

$$n \geq \frac{\log\left(1-\frac{g(1-\delta)}{(L-1)\delta}\right)}{\log(\delta)} \tag{6.5}$$

That is, A's threat, n, must be large enough to deter B from small defecting first. We can also characterize when no such n exists. The limit of the above as $n \to \infty$ gives us the condition:

$$\delta < \frac{g}{L-1+g} \tag{6.6}$$

□

This implies that if δ is relatively small compared to the given ratio, the shadow of future punishments will not be an effective deterrence to B.

However, we note that Equation 6.5 is similar to Equation 6.2, and are related in the following way:

$$\frac{g(1-\delta)}{(L-1)\delta} < \frac{(1-\delta)}{(L-1)\delta} \leq (1-\delta^n) \tag{6.7}$$

This is true because of the assumption that $g < 1$. This leads to the following:

Theorem 6.3.3. *If effective punishments exist to deter large defection, then effective punishments also exist for small defections.*

Intuitively, if A can effectively dissuade B from large defection with the threat of mutual defection, then it can also deter the temptation for B to pursue the lesser payoff of g. Therefore our condition for cooperation by A is $\delta > \frac{1}{L}$ because this implies that future cooperation is worth more than A risking any kind of defection,

and being punished by B in the future. Further, if $\delta > \frac{1}{L}$, then B is dissuaded from large defecting, but not necessarily from small defecting.

6.3.4 Credible Punishment

We can express whether the punishment is credible by examining whether carrying out the punishment increases A's payoff over simply allowing B to small defect. That is, is it better for A to tolerate a small defection from B in all rounds, or is it better off attempting to punish a small defection with n rounds of large defection?

Theorem 6.3.4. *No credible punishments against small defect exist if* $\delta < 1 - \frac{\lambda L}{L-1}$

Proof. We compare the payoff for A punishing a low defection for n rounds against A simply tolerating a small defect in all rounds.

$$\begin{aligned} -\lambda L \sum_{i=0}^{\infty} &\leq -\lambda L + (1-L)\sum_{i=1}^{n} \delta^i \\ \frac{-\lambda L}{1-\delta} &\leq -\lambda L + \frac{(1-L)\delta(1-\delta^n)}{1-\delta} \end{aligned} \tag{6.8}$$

This simplifies to:

$$(1-\delta^n) \leq \frac{\lambda L}{L-1} \tag{6.9}$$

$$n \leq \frac{\log\left(1 - \frac{\lambda L}{L-1}\right)}{\log(\delta)} \tag{6.10}$$

Equation 6.9 implies that if n is small enough, then punishing B is preferable to tolerating small defections. However such an n does not always exist. If the smallest

possible n does not provide a credible punishment, then none will. For $n = 1$ (the smallest possible punishment), if:

$$\delta < 1 - \frac{\lambda L}{L - 1} \tag{6.11}$$

then no possible n gives a credible threat. □

That is if δ is small compared to the given fraction, A would rather take a small loss for the duration of the game than risk the immediate large loss from mutual large defects.

We finally examine the credibility of punishments of large defections.

Theorem 6.3.5. *Punishment of large defections is always credible.*

Proof.

$$-L \sum i = 0^{\infty} \leq -L + (1 - L) \sum_{i=1}^{n} \delta^i \tag{6.12}$$

$$-L + L(1 - \delta) \leq \frac{(1 - L)\delta(1 - \delta^n)}{1 - \delta} \tag{6.13}$$

$$\frac{L}{L - 1} \geq (1 - \delta^n) \tag{6.14}$$

Because $\frac{L}{L-1} > 1$ is always true, and $0 < 1 - \delta < 1 - \delta^n < 1$ Equation 6.14 is always true for any n, and that the threat to punish against a large defection is always credible.

□

6.3.5 Credible or Effective, but not both

A final possibility is that for some n punishment is credible but ineffective, or that the punishment is effective but not credible.

Theorem 6.3.6. *Effective and Credible punishments are mutually exclusive if* $\delta < \frac{1}{\lambda L+1}$

Proof. If the condition in equation 6.9 holds but 6.2 does not, then:

$$\frac{\lambda L}{L-1} < \frac{(1-\delta)}{(L-1)\delta} \tag{6.15}$$

An equivalent condition is reached if the condition in equation 6.9 does not hold but 6.2

Simplification of the above gives us the final condition: $\delta < \frac{1}{\lambda L+1}$. □

6.3.6 Responsibility for d

Above we have given conditions under which various equilibria exist in the APD. They are:

- $\delta < \frac{1}{L}$: No punishments are effective, because the temptation to defect is too large. Mutual Defection in all rounds.
- $\delta > \frac{1}{L}$ and $\delta < 1 - \frac{\lambda L}{L-1}$: Punishments are effective but not credible. A cooperates while B small defects in all rounds.
- $\delta > \frac{1}{L}$, $\delta > 1 - \frac{\lambda L}{L-1}$, and $\delta < \frac{1}{\lambda L+1}$ Effective punishments exist, and credible punishments exist, but they are mutually exclusive. A cooperates while B small defects in all rounds.

- $\delta > \frac{1}{L}$, $\delta > 1 - \frac{l}{L-1}$, and $\delta > \frac{\lambda L}{L+1}$: Effective and credible punishments against small defection exist. Mutual cooperation in all rounds.

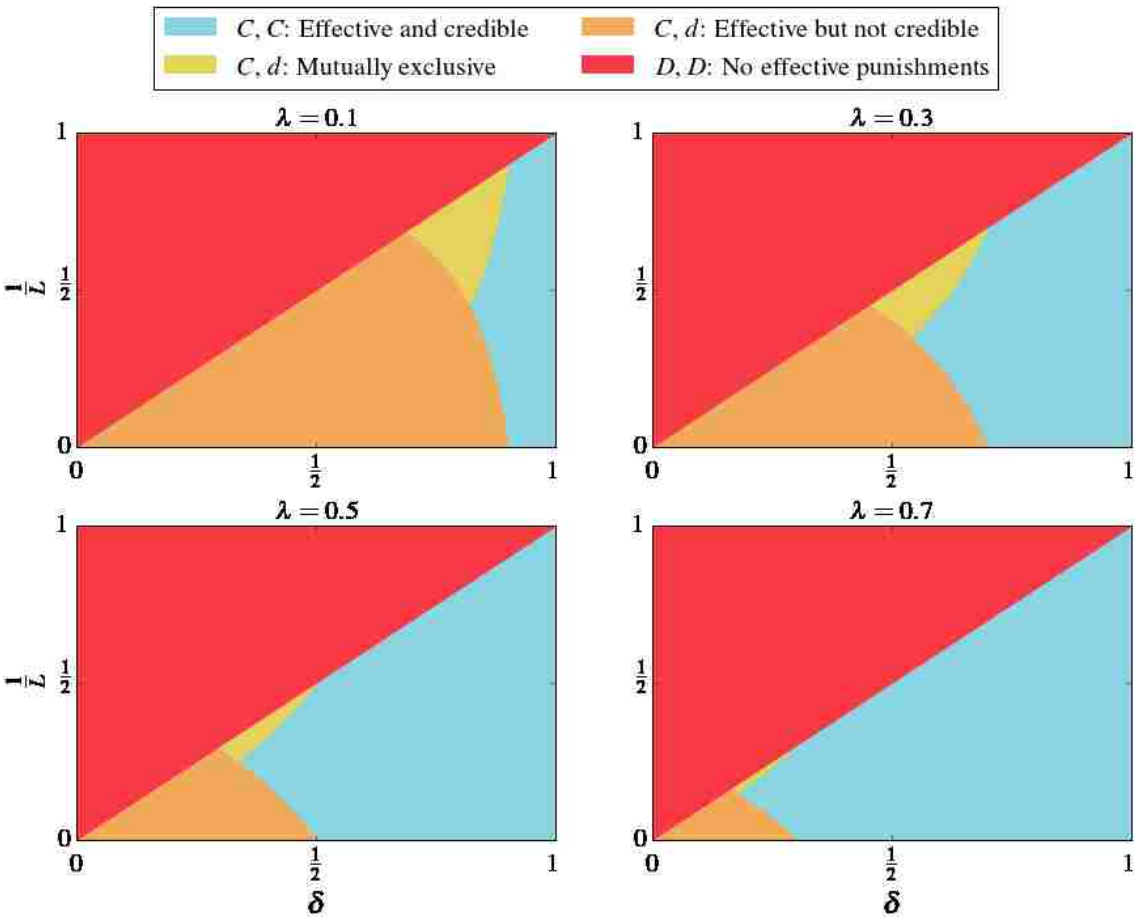

Figure 6.2: The equilbrium space of the asymmetric prisoner's dilemma for four example values of λ.

We note that the conditions for equilibrium outlined above depend only on three variables: L, δ, and λ. g only appeared in the conditions for effective punishments of small defection, and we saw that this condition is subsumed under the conditions for effective punishments for large defection, and so does not influence the equilibrium. Moreover, if we consider $\frac{1}{L}$ instead of L, all three of these parameters are in the

range of 0 to 1. We illustrate the space of parameters and the resulting equilibria in Figure 6.2.

As can be seen in this figure, as λ decreases it is more difficult for A to have credible and effective punishments to enforce mutual cooperation. This may be the case with the US and North Korea in the case of the Sony attack. The harm caused by the attack was small, and no effective proportional punishments existed. The US then had to result to the out of domain response of economic sanctions.

6.4 The Attribution Game

In this section we explore a final modified version of the Prisoner's Dilemma(PD) that further illustrates the relevance of game theoretic models to cyber conflict. We call this modified game the Attribution Game(AG).

6.4.1 Attribution Game Description

The AG consists of two players (A and B) playing a single round two component game. First the players play a standard simultaneous prisoner's dilemma, with each player choosing to cooperate, C, or defect D against the other, without knowing the other players move.

For this game, we move back to the standard prisoner's dilemma payoffs. That is players receive reward R for mutual cooperation, Temptation T for defecting when the other player cooperates, suckers payoff S if they cooperate while their opponent defections, and punishment P for mutual defection, with $T > R > P > S$. The normal form of the payoff matrix can be seen in table 6.3.

A second component of the game is played only if one player cooperates (the

Table 6.3: The standard prisoner's dilemma between two players A and B.

		B's Move	
		Cooperate	**Defect**
A's Move	**Cooperate**	R, R	S, T
	Defect	T, S	P, P

victim) while the other defects (the attacker). The victim is given the opportunity to punish the attacker. This punishment takes the form of an additional defection, in which the attacker is contrite with probability α, that is they accept the punishment and do not retaliate. With probability $1-\alpha$ the attacker is not contrite and responds with a punishment of their own. To allow for asymmetry in the game we denote the probability that B is contrite to A's punishment as α and the probability that A is contrite to B's punishment as β. We assume that each player knows the payoffs and the opponent's values of α and β respectively.

Substantively, we view contrition as a function of the victim's ability to publicly attribute the defection to the attacker. We view contrition as a combination of three factors, specifically, the value represents: technical evidence, responsibility, and reputation. We discuss this more in section 6.5. The payoffs for this game can be seen in table 6.4.

6.4.2 Game Analysis

In this section, we determine all possible equilibria given the attribution ability, α and β, of each player. Despite the relatively simple expansion, many of the equilibria

Table 6.4: Payoff matrix for the expanded attribution game. A's moves are all represented in the rows of the matrix (vertical text), and B's moves are all represented in the columns of the matrix (horizontal text).

B's First Move

A's First Move		**Cooperate**		**Defect**	
	Cooperate	R, R		**A's Second Move**	**Don't Punish**	S, T
				Punish	$S + P + \alpha(T - P)$, $T + P + \alpha(S - P)$
	Defect	**B's Second Move**		P,P	
		Don't Punish	**Punish**		
		T, S	$T + P + \beta(S - P)$, $S + P + \beta(T - P)$		

are non-trivial, and in some cases no pure strategy equilibria exist.

Punishing Defection

We start by determining if, when a player is attacked, if she should punish her opponent. Without loss of generality, we examine the case from A's perspective, noting that B's perspective is symmetric.

Theorem 6.4.1. *A should risk punishment if* $\alpha \geq \frac{-P}{T-P}$

Proof.

$$S \leq \alpha(S + T) + (1 - \alpha)(S + P) \tag{6.16}$$

$$\alpha \geq \frac{-P}{T - P} \tag{6.17}$$

□

This condition holds when $P > 0$, that is, if the payoff for of mutual defection is positive, players always have an incentive to punish their opponents. If the condition in 6.17 doesn't hold for both players, then the game reverts to the normal PD where the only equilibrium is mutual defection. If either player fails to meet the criteria, then mutual cooperation is no longer an equilibria.

Risking Defection

If the condition in equation 6.17 is met for both players then each player may choose to risk defection, and subsequent punishment. Once again, we examine the possibilities from A's perspective.

Theorem 6.4.2. *A should defect if* $\beta < \frac{T+P-R}{P-S}$.

Proof.

$$R \leq \beta(T+S) + (1-\beta)(T+P) \tag{6.18}$$

$$\beta \leq \frac{T+P-R}{P-S} \tag{6.19}$$

□

Equation 6.19 indicates that if the opponent's ability to attribute is small enough, then it is rational to defect, and risk being punished.

Pre-emptive Defection

A player who believes she will be the victim of a defection (conditions in equations 6.17 and 6.19 are met), may decide to cooperate anyway, because punishing their attacker may be preferable to mutual defection.

Theorem 6.4.3. *Punishment is preferable to mutual defection if* $\alpha \geq \frac{-S}{T-P}$

Proof.

$$P \leq \alpha(S+T) + (1-\alpha)(S+P) \tag{6.20}$$

$$\alpha \geq \frac{-S}{T-P} \tag{6.21}$$

□

This condition is slightly stronger than the one given in 6.17 as $\frac{-P}{T-P} < \frac{-S}{T-P}$ by the assumption in the PD that $P > S$.

However, the condition in 6.19 has no natural ordering with the other two conditions, and each of the following three orderings are possible (depending on the values of T, R, P, and S).

$$\frac{-P}{T-P} < \frac{-S}{T-P} \leq \frac{T+P-R}{P-S} \tag{6.22}$$

$$\frac{-P}{T-P} \leq \frac{T+P-R}{P-S} < \frac{-S}{T-P} \tag{6.23}$$

$$\frac{T+P-R}{P-S} \leq \frac{-P}{T-P} < \frac{-S}{T-P} \tag{6.24}$$

α and β's values within these three possible orderings of thresholds will determine the equilibrium.

6.4.3 Equilibria

The above conditions allow us to calculate the exact conditions for various equilibria. As we noted above, if 6.17 is not true for either player the game reverts to a prisoner's dilemma, and the only equilibrium is for both players to defect. However, if we assume that 6.17 is true for both players, a new finite game emerges, the payoff matrix can be seen in 6.5.

Table 6.5: The extended prisoner's dilemma when $\alpha > \frac{-P}{T-P}$ and $\beta > \frac{-P}{T-P}$.

		B's Move	
		Cooperate and Punish	**Defect**
A's Move	**Cooperate and Punish**	R, R	$S + P + \alpha(T - P)$, $T + P + \alpha(S - P)$
	Defect	$T + P + \beta(S - P)$, $S + P + \beta(T - P)$	P, P

Equations 6.19 and 6.21 form the conditions under which players will change their strategy in this new game. Given that there are only 16 possible ways for each of the conditions to be met we can check them exhaustively. We find that this produces in six possible universes.

- Mutual Cooperation is the sole equilibria.
- Mutual Defection is the sole equilibria.
- Mutual Cooperation and Defection are both equilibria.
- One player punishing another is an equilibria.
- Each player punishing the other is an equilibria.
- No pure equilibria strategies exist.

We can see for what orderings given in 6.24 and values of α and β each equilibria exist in Figure 6.3.

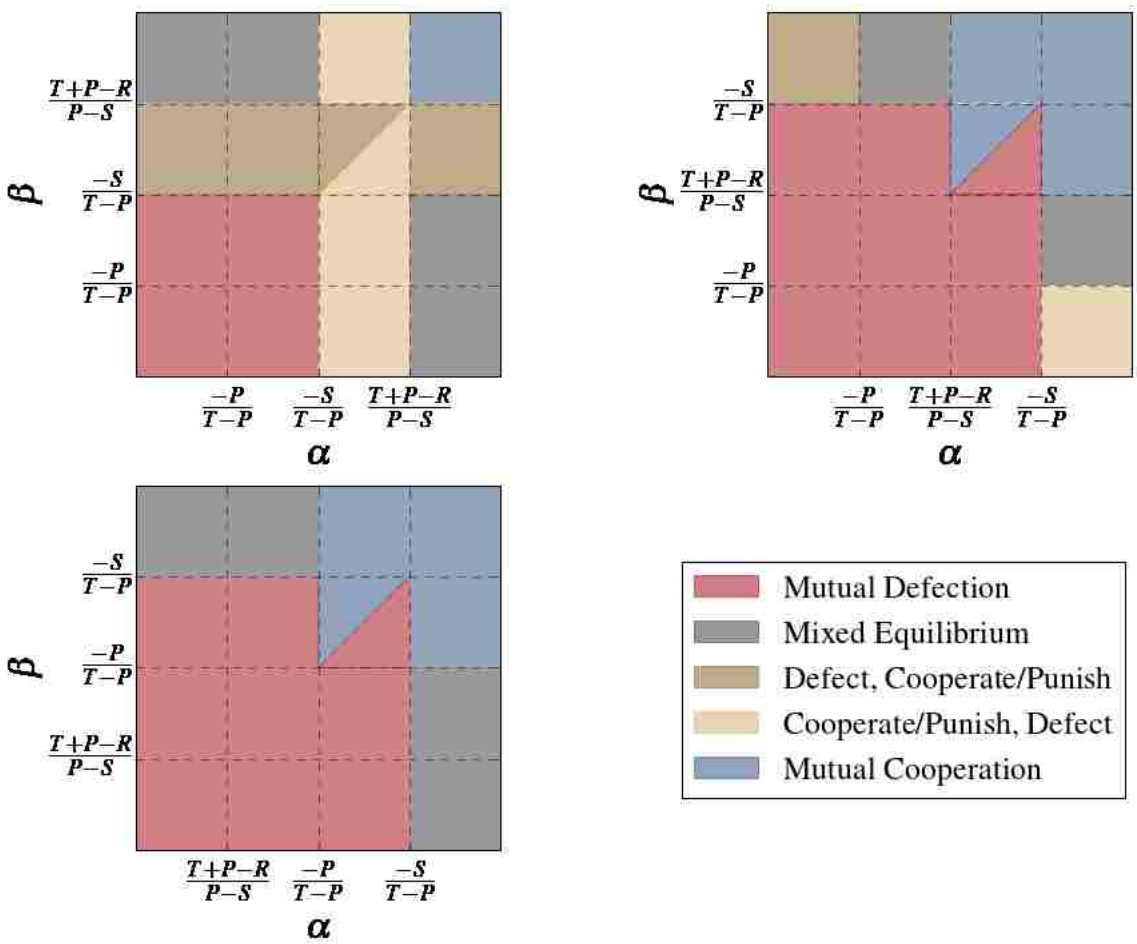

Figure 6.3: All possible equilibria, for possible orderings of the thresholds and possible values of α and β. For each colored area above, if α and β fall into that area, then that area represents the equilibria of that particular parameterization of the game. If two colors are in that area (separated by a diagonal) then two equilibria are possible.

Mutual cooperation emerges as an equilibrium if $\alpha > \frac{T+P-R}{P-S}$ and $\beta > \frac{T+P-R}{P-S}$. Mutual defection is an equilibria if $\alpha < \frac{-S}{T-P}$ and $\beta < \frac{-S}{T-P}$. We note that these conditions are *not* mutually exclusive, and so there do exist conditions under which both are equilibria. In this case AG takes on the characteristics of a coordination game [54].

When a small imbalance in attribution ability exists, then punishment becomes an equilibrium. In particular, if both players satisfy $\{\alpha, \beta\} > \frac{-S}{T-P}$, but only one player's ability to attribute is greater than $\frac{T+P-R}{P-S}$, then defection and cooperate/punishment is an equilibrium. Interestingly, the player with the lower probability of attribution will be the one punishing the other player (Figure 6.3 top two panels).

Finally, if there is a large difference in attribution ability, $\alpha > \frac{-S}{T-P}$ and $\alpha > \frac{T+P-R}{P-S}$ while $\beta < \frac{-S}{T-P}$ and $\beta < \frac{T+P-R}{P-S}$, then no pure strategy equilibria exist. In this case, a mixed strategy equilibrium exists, with each player playing cooperate with probability

$$q = \frac{S - \beta(T-P)}{P - R + S + T\alpha(S-P) + \beta(T-P)} \tag{6.25}$$

6.5 Discussion

We suggest that the models presented in this chapter are relevant to cyber conflict between the United States and China. For example, the U.S. was recently the victim of cyber attacks that stole personal information about nearly 22M federal employees. Numerous U.S. news outlets reported that the attack originated in China [19, 199], and were sophisticated enough to the be the work of a nation state. Despite what seems to be reliable technical attribution of the attack, the U.S. has refrained from blaming China publicly. However, China eventually agreed to arrest those it claims are responsible, but it is unclear what punishment they will face [200].

The models given in sections 6.2 and 6.3 provide clues for interpreting this action. If the U.S. (A in our games) is unsure of the exact relationship between the attackers and the country they are from, then it is prudent not to retaliate. The waters were further muddied in this case when China (B) offered to arrest the hackers

(C) responsible for the OPM breach, to make it appear as if they did not directly benefit from the attacks. It may also be the case that even if the relationship were known by the U.S., the lack of a proportionate response could explain the U.S.'s lack of response, the exact behavior we see in the Asymmetric Prisoner's dilemma (section 6.3).

This event resembles those reported by Mandiant, a U.S. cyber security company, in 2013 [36]. This report provided details on a Chinese cyber espionage unit, and their activities over the course of roughly a decade. During this time, the unit was responsible for numerous acts of industrial espionage. In this case, the U.S. response was also muted. While the attribution presented in the Mandiant report was largely accepted by the security community and public [257], the only official action by the U.S. was indicting five Chinese nationals believed to be responsible for the attack [211]. In response to the indictments, China denied involvement in the alleged crimes and claimed the U.S. had ulterior motives for the indictments [18]. However, in 2015 was reported that the Chinese military had scaled back industrial cyber-espionage in response to these indictments [201]. Documents later released by Edward Snowden indicating that the U.S. National Security Agency (NSA) does conduct operations against Chinese corporations such as Huawei and China Telecom [238]. However, the NSA claims that, in contrast to the Chinese attacks, these efforts against foreign corporations are not made to benefit U.S. corporations [71].

This particular situations may be well modeled by the game presented in section 6.4. Here the U.S. and China are playing the Attribution Game (AG), where their particular ability to attribute, α and β, are low enough to result in mutual industrial espionage (mutual defection). Increasing the U.S. attribution ability might not be sufficient to bring about mutual cooperation, however, as in the AG only high levels of technical attribution by both players lead to mutual cooperation.

It is often assumed that the only requirement for a nation to justify retaliating

against an attack by an adversary is sufficient evidence to prove that the adversary conducted an attack (technical attribution). Many game theoretic models of conflict assume that each player has full knowledge of the action of the others. However, in a growing number of conflicts, including terrorism and attacks in the cyber domain, it is difficult to determine the provenance of an attack, or which actors can reasonably held responsible.

There are some models where this assumption is relaxed [88, 172]. In these models, it is assumed that attribution is simply the probability of having evidence that an attack can be attributed to an opponent. Inevitably, this leads to the conclusion that the best strategy for players is to increase their technical ability to trace attacks back to their source, or at the very least do so in proportion to the severity of the attack [172]. Here we conceptualize attribution ability, not as the technical capability to perceive the origin of an attack — although this capability is related — but rather as ability and willingness to publicly accuse another actor of an attack and provide sufficient evidence to convince the international community that punishment is justified, rather than an unjustified violation of international norms. We find that under these circumstances the attribution abilities of *both* actors are critical in enabling cooperative equilibria. This implies that increasing the attribution ability of the U.S.'s adversaries, such as China, could actually increase the chance for cyberpeace.

6.5.1 Attribution in the Attribution Game

An actors ability to punish an opponent for defecting against it without the opponent retaliating, then, is a function of the actors attribution ability. Any public punishment implies an evidentiary appeal (either explicit or implied) to the international community to verify that the target of the punishment was the original aggressor.

As we indicated in section 6.4, this is a function of technical evidence, evidence of responsibility, and ongoing reputation. A successful public justification for an attack results in contrition by the punished player – that is, the punished player does not treat the punishment as a defection. If a player punishes but is unable to justify its action, the punished player perceives the punishment as a defection in response to which the punished player sequentially defects. We parameterize the probability of an attacker accepting punishment (and not retaliating) as a probability $\{\alpha, \beta\} \in [0, 1]$, for actors A and B respectively in the attribution game. This probability is directly related to attribution ability.

For example, we can apply this understanding of attribution to explain the retributive dynamics of U.S. - China cyber relations. In particular, if the U.S. has significantly more advanced attribution ability [45] (high α) than China does (low β), our model predicts possible equilibrium where both actors defect without attempting to punish the other. From figure 6.3, we can surmise that in this case the U.S. has enough attribution ability for credible and effective punishments, but China does not. This suggests a situation where the U.S. and China are willing to engage in repeated cyber defections with neither willing to risk public acknowledgment and punishment of these defections, i.e. (D,D) no punishments. China recently claimed to arrest those responsible for Office of Personnel Management attack [200]. However, it is unclear whether those arrested were actually responsible, or, if they are, what punishment they face. Moreover, such arrests could be intended to lower the U.S.s certainty that the attack was government sponsored. The inability of China to correctly attribute the attacks, meant that the U.S. was able to covertly defect (mutual defection) without public accusations. Only when documents were leaked by Edward Snowden was it known that the U.S. was engaging in this activity [238].

Counter-intuitively, if Chinas ability to attribute cyber attacks was enhanced, it might push the game into an equilibrium of mutual cooperation. If both the U.S. and

China had credible and effective punishment, this would prevent either side from risking defections. To summarize, this model shows how critical uncertainties in cyber attribution could impede cooperative equilibria that achieved in traditional/kinetic conflicts, where attribution is unambiguous. The more that attribution in cyber-conflict improves and resemble attribution in kinetic attack, the more we can expect traditional mechanisms of mutual deterrence will be effective.

At the same time that we may strive to enable cooperation, we must acknowledge that the traditionally more transparent kinetic and nuclear conflicts have begun to resemble cyber. The rise of global terrorism and well-resourced non-state actors has multiplied the possible sources of sophisticated attacks. These entities tend to be embedded within civilian communities, and to operate in geographies over which state actors claim jurisdiction, complicating issues of attribution. This suggests that we should begin to adapt our thinking on kinetic conflicts to address this ambiguity. Our model maybe of use in this process.

As we outline above, our models (particularly the RG) can be applied to more traditional conflicts as well. Conflict between Palestine and Israel has been rife with both covert and overt actions over more than 60 years[6]. Groups such as the Palestinian Islamic Jihad have conducted numerous attacks against Israel in the past 30 years [121], with each attack provoking a varied response from Israel. The question of attribution here is not necessarily one of technical ability as it is in the U.S./China cyber conflict case, but whether the ruling party in Palestine, currently Hamas, can be held responsible for the actions of these terrorist groups, that is, how strong are the values of λ_A and λ_B. If no significant affiliation can be discerned between terrorist groups and the governments for which they purport to fight, then punishing such a group could lead to condemnation from the international community. The varied political climate of the region likely results in different estimates of these parameters

[6]Thanks to Robert Axelrod for pointing us to this particular example

for Israel and accounts for the variety of strategies we see used in the region when responding to new attacks.

6.5.2 Repeated Play and Reputation in the Attribution Game

The analysis of the AG presented above focused on a single round of play, and even in this simple form produced insight into cyber conflict. Our definition of the probability parameter α suggests that exploring multiple round versions of the AG is a fruitful avenue for further work. For example, we could both endogenize and decompose α.

At the beginning of the game α_r might represent the starting reputation of the player, e.g. the ability to attribute an attack to the other player in such a way that a punishment is justified. Defections might provide evidence to strengthen the attribution and the probability that a player can carry out a justified punishment increases. That is $\alpha = \alpha_r + e(k)$, where k is the number of defections and α_e is the evidence gained with each defection[7].

When a punishment is attempted the player would then have to reveal evidence of the defections to be able to retaliate. This would allow the punished player to alter its attack strategies in the future, making the gained evidence useless in future rounds. Whether the punishment is justified or not could alter α_r. Justified punishments can increase a countries reputation $\alpha_r \leftarrow \alpha_r + r$, and if the punishment is unjustified α_r could decrease ($\alpha_r \leftarrow \alpha_r - r$)[8]. This would allow countries to build reputation through justified punishments, and lose reputation through unjustified ones. Reputable countries would be able to respond to nearly all defections

[7] $e(k)$ would be some increasing function of k and α such that $0 < \alpha < 1, \forall k$.

[8] We use a linear increase, but more likely r could be a percentage increase towards 1 or 0.

with punishments, and therefore deter future defections, while disreputable countries would have to build large amounts of evidence to respond with punishments.

We might expect different countries to have different reputation values α_r and their reputations might grow and different rates. Additionally, their ability to collect evidence of attacks might differ. So some players might gain reputation and evidence faster than others, gaining them an advantage.

Moreover, α could increase or decrease based on a cost played by a player. That is a defecting player could decrease the opponents α by investing in technologies to 'cover their tracks'. Similarly, a player could invest in investigation techniques to increase the α and their chance of presenting convincing evidence making payoffs justifiable. This could introduce an interesting dilemma for player's on where to invest resources, in hiding their own attacks or investigating their opponents. Appropriately modeling the return on investment in these cases could lead to different outcomes and model a number if situations.

6.6 Conclusion

In this chapter we explored a number of game theoretic models for cyber conflict. We have discussed three unique aspects of cyber conflict that can lead to some surprising results. Imperfect knowledge of the relationship between attackers and nation states can make it difficult for nations to determine appropriate actions. A lack of proportionate response by one player, can lead to situations where it is rational for a player to accept continued defection by another. Finally, raising the attribution ability of an adversary can lead to a higher likelihood of mutual cooperation in many situations.

Chapter 7

Conclusions

As stated in chapter 1, this dissertation argues that achieving computer security requires both rigorous empirical measurement and data-driven and abstract models to understand cybersecurity phenomena and to determine which defenses and interventions will be most effective. We used this perspective to show that such analysis and modeling can reveal surprising results about a variety of different cybersecurity problems. Here we review those findings, and provide some final remarks.

7.1 Summary of Findings

- **Categorization of Security Interventions**: In Chapter 2 we gave a general overview of the computer security research. While this was meant to frame the need for more rigorous modeling and analysis, the presented categorization of security research is novel. In particular, it may serve in the future as a theoretical framework for understanding what defenses are effective against which attacks.

- **There is no evidence for reported increases in data breaches size or**

frequency: Next, we presented a rigorous analysis of the most complete known dataset of public data breaches. We found that despite media attention, data breaches have not increased in size or frequency in the last ten years. Moreover, the specific distributions in the size and frequency in data breaches provides clues for the mechanisms that may generate them. Finally, we provide some predictions about the likelihood of future breaches, and their cost.

- **Botnet takedowns have a geographically varied effect on concentrations of spam sending IP addresses**: In Chapter 4, we presented an analysis of a spam dataset spanning 10 years, 60 countries, 260 ISPs, 127 million IP addresses, and 440 Billion messages. We identified geography, national economics, and internet connectivity as external factors that are correlated with local spam concentrations. Moreover, we found that botnet takedowns, a popular approach to managing spam, has been largely ineffective over the last ten years, and that other interventions such as real-time adaptive filter, and value chain interventions are more likely responsible for global declines in spam.

- **Depreferencing is an alternative for controlling user exposure to malicious websites in search results**: In Chapter 5 we constructed an abstract model of web infections and control. We used the model to show that website depreferencing is a reasonable alternative to blacklisting malicious websites. Moreover, depreferencing allows search providers to balance user exposure to malware with harm to the traffic lost due to websites being incorrectly labeled as infected. Our approach shows how abstract models can determine outcomes through simulation without the need for costly experiments, and points to potential problems with experiments.

- **Attribution in cyber conflict is more than a technical challenge, it is a strategic question**: In our final chapter, we examined a number of game theoretic models of cyber conflict. These models demonstrate that uncertainty

about the relationship between attack and adversary can lead to unnecessary conflict. Even when the relationship is known, if no proportional responses to attacks exist it may be rational to accept small attacks rather than risk escalation. Finally if there is a large asymmetry in the ability of nations to attribute attacks, counter-intuitive behavior, such as the allowing attacks to go unchecked, becomes rational.

7.2 Future Work

The research presented in this dissertation provides numerous opportunities for future work, both as extensions to the current research, and new research that adopts the approach taken here. Chapter 3 indicated that there was no overall trend in data breach size or frequency, but did not examine trends within different organizations. Recent work suggests that breaches among healthcare organizations are increasing [231], and though we caution drawing too many conclusions from industry reports, our methodology could generally be applied to specific industries. Further we could examine if there are risk factors for different organizations experiencing a breach. Recent work has shown that it is possible to *predict* breaches [175], but gives little indication about which organizational features increase the risk of a breach. Finally, the cost models given in 3 are rough estimates at best. More precise data about the cost of breaches could help develop better estimates of the risk organizations face when storing personal information.

Our examination of global spam levels provided insight into which interventions have been effective in the past decade. Our approach could easily be applied to other malicious activity and other interventions. We could also use current or future data to examine the effect of interventions such as national botnet clean-up programs such as Germany's BotFrei [30]. Our work examined concentrations of spam, but

did not ultimately link spam to specific costs to countries, ISPs, organizations, or individuals. We could leverage information on messages sent by ISPs to try to identify the cost of individual messages to recipients to understand the financial effect of spam interventions. Finally, we found that some countries and ISPs experience increases in spam concentrations after takedowns. Future work on identifying countries at risk for these increases would give policy makers guidance as to where efforts at cyber-capacity building would be most effective.

Our game theoretic approach to cyber conflict developed a number of interesting conclusions. We believe these models could be further developed to apply to other types of conflicts such as terrorism. We introduced unique dynamic, shared payoffs in the Responsibility Game in chapter 6. Studying this in a general way could lead to new insights in why agents organize in social systems.

Finally, our general approach could be used to begin to understand long standing security questions. For example, given sufficient data, we may be able to correlate an organization security practices, such as their ability to keep critical software up to date, with outcomes such as botnet infections, or data breaches. It has been hypothesized that software monoculture, the phenomena where the majority of systems all use the same software, e.g. Windows, allows major vulnerabilities to be widely exploited and put the cyberspace at unnecessary risk [92, 206]. However, the alternative, software diversity, could increase the attack surface of a system [94]. A data driven approach could help understand the likely complexities of this problem. Additionally, appropriate measurement and modeling could be applied to identify the effect of data locality on the performance and growth of the Internet or the effect of censorship on the flow of information and ultimately political action.

7.3 Final Remarks

With unprecedented numbers of people now connected to and depending on the Internet (three billion in 2014 [275]), it is imperative that we understand and mitigate global cybersecurity threats. Further, we need to understand regional variations, and why some parts of the world and some corners of the Internet are disproportionately affected. This is especially important as users begin to understand the dangers of cyber insecurity and policymakers begin to act to try to protect the public.

We believe that models like the ones we present here can provide guidance in these matters. These models are not without challenges. As we have shown, heavy-tailed distributions make it difficult to distinguish trends and the effect of interventions in data breaches, spam, and the spread of malware. Simply comparing one time period's measurements another's is unlikely to provide an accurate or meaningful picture.

Collecting and analyzing appropriate data will continue to be a challenge. For example, though we did our best to analyze the most complete public dataset of breaches in chapter 3, it is likely that many breaches go unreported despite breach reporting laws. Similarly, our spam data was collected from a single spam trap, and may have only observed a portion of global spam dynamics, though we did make efforts to validate our data with others. If we wish to truly understand security using data-driven modeling, a more open sharing of security data will be necessary. National initiatives such as the Federal Cybersecurity Information Sharing Act are a preliminary step towards this goal [33].

Data alone will not provide the solution though. Proper analysis and methods must be used that plausibly link attacks and defenses to outcomes. It is popular today to frame cybersecurity outcomes in terms of risk analysis and management, and we take the view that this is ultimately the correct approach. For example,

the U.S. National Institute of Standards (NIST) has developed and promulgated its cybersecurity framework, which is based almost entirely on the concept of risk assessment [212].

By building and analyzing plausible models of the link between attacks, defenses and outcomes, we will be better able understand where policy makers and technological innovators should focus their efforts for reducing cyber-insecurity, while doing their best to balance positive effects with negative externalities. Abstract models like those presented in chapters 5 and 6 can help researchers avoid the large-scale (and potentially expensive) experiments needed to test interventions in the field.

In this dissertation we have demonstrated that surprising conclusions can be drawn about a wide variety of security phenomena when appropriate analysis and modeling are applied. As the scope of digital life increases so will the need for knowledge about the best approaches to protect the public. This growth has already caused the scope of cyber security problems to grow beyond the ability of any one security practitioner from being able to grasp all the relevant details associated with global problems [93]. The solution must be further application of models similar to those we have presented in this dissertation.

Appendices

Appendix A

A Strategies within the Responsibility Game

A.1 Phase strategies

In the equilibrium analysis in section B we employ a set of strategies known as punishment phase strategies, adapted from common strategy profiles from the literature on infinitely repeated games [80]. In these strategies play is comprised of two types of play: normal play and punishment phase play. During normal play, each player takes an action prescribed by the proposed equilibrium. For the purposes of our analysis, during any round of normal play action set α_{peace}, detailed in A.2.1, is played.

If any player deviates from normal play, the game enters a punishment phase beginning with the immediate next move (whether this is the next round or simply the subsequent move, as in when punishment on C occurs later in the same round because of the ordering of actions) in which the deviating player is the target. All players then play according to one of the punishment action sets for deviating player X, which yield $U_X(\alpha_X) < U_X(\alpha_{peace})$ during the subsequent punishment phase, which

lasts for n rounds, where $n \in [1, \infty)$. A variety of different action sets, enumerated in A.2.3, can be played during punishment phases.

In the limits, then $n = 1$, the punishment phase amounts to a three player version of a single round tit-for-tat strategy, and as $n \to \infty$ the punishment approaches a grim trigger.

A.2 Action Sets

The strategies used during normal play and punishment phases of the responsibility game invoke specific action sets to indicate the behavior of the players in different conditions. An action set is the set of actions carried out by each player in each of the three components of the game. We introduce the reader to these below.

First we introduce some notation. Let α_i indicate the action set in a given round i, such that $\alpha_i = \{\alpha_i^A, \alpha_i^{B_{PD}}, \alpha_i^{B_P}, \alpha_i^C\}$, where α_i^A is player A's action in round i, $\alpha_i^{B_{PD}}$ is player B's action in its Prisoner's Dilemma with A in round i, $\alpha_i^{B_P}$ is player B's choice wither to penalize C in round i, and α_i^C is player C's Attack/Not Attack action against A. We denote the utility of a player X for an action set played in a single round of the RG as $U_X(\alpha_i)$.

A.2.1 Peace

As we will discuss in more detail below, the Folk Theorems indicate that the range of outcomes in which payoff vectors exceed minimax payoffs can be supported as subgame perfect equilibria in the infinitely repeated Responsibility Game [91]. As in Fearon and Laitin [89], we are interested in explaining the conditions under which it is possible for a specific and substantively valuable peaceful, cooperative outcome is

a supportable equilibrium. In particular, we focus on the action set we call "peace", denoted α_{peace} in which players A and B cooperate with each other in their Prisoner's Dilemma, B does not sanction C, and C does not attack A. Thus, the action set for a round of peaceful play is, $\alpha_{peace} = \{C, C, NP, NA\}$.

Payoffs from the game defined above are scaled such that the single round payoffs for each player when $\alpha_i = \alpha_{peace}$, are zero as are given by the following utility functions:

$$U_A(\alpha_{peace}) = 0 \tag{A.1}$$

$$U_B(\alpha_{peace}) = 0 \tag{A.2}$$

$$U_C(\alpha_{peace}) = \lambda_C 0 + 0 = 0 \tag{A.3}$$

A.2.2 War

The converse strategy in the RG, in which all players are best responding to their immediate interests in the stage game yields an action set we call War, $\alpha_{war} = \{D, D, NP, A\}$.

The single round payoffs for each player when $\alpha_i = \alpha_{war}$ are given by the following utility functions:

$$U_A(\alpha_{war}) = G - L \tag{A.4}$$

$$U_B(\alpha_{war}) = G - L + \lambda_B \tau \tag{A.5}$$

$$U_C(\alpha_{war}) = \tau + \lambda_C(-(G - L))\lambda_C G \tag{A.6}$$

Later we will use descent into war is used as an implicit threat to enforce more cooperative equilibria.

A.2.3 Punishment actions

To enforce a peaceful equilibrium, we will rely on a set of punishments for deviation from peace. Different action sets may be used to punish each player, as any action set that lowers a players utility can be considered a punishment. Below we detail the range of single round action sets that can be played as part of the process of punishing a defecting player. Here we denote an action set α_X as the set of actions played by all three players to punish player X, with $\alpha_X^Y \in \alpha_X$ denoting any player Y's action prescribed by α_X.

The harshest punishment actions that can be enforced on an unwilling player are the minimax punishments, where the minimax value for a player is given by $\overline{v_i} = \min_{\alpha^{-i}} \max_{\alpha^i} v_i(\alpha^i, \alpha^{-i})$

$$\alpha_{minimax_A} = \{D, D, NP, A\} \tag{A.7}$$

$$\alpha_{mimimax_B} = \{D, D, NP, NA\} \tag{A.8}$$

$$\alpha_{minimax_C} = \{D, C, P, A\} \tag{A.9}$$

The payoffs for these strategies, which we have discussed above, are reproduced here for convenience:

$$U_A(\alpha_{minimax_A}) = \overline{v_A} = G - L - d_{CA} \tag{A.10}$$

$$U_B(\alpha_{minimax_B}) = \overline{v_B} = G - L \tag{A.11}$$

$$U_C(\alpha_{minimax_C}) = \overline{v_C} = \tau - d_{BC} + \lambda_C(0 - G) \tag{A.12}$$

The Folk Theorems indicate that any outcome that has an average payoff vector $v* \geq \overline{v}$ can be sustained as a subgame perfect Nash equilibria [91].

In the following sections we explore a larger set of punishment action sets which we find more theoretically and substantively meaningful than the minimax action set. Any action set α_X for which $U_X(\alpha_X) < U_X(\alpha_{peace})$ in a single round of play can be used as a punishment action set on player X. Because $U_X(\alpha_{peace}) = 0$ for all X, any action set that yields a negative payoff for a given player can be used as a punishment action set for that player. The viable pure strategy punishment action sets are enumerated for each player in the tables below.

Action Set	$U_A(\alpha_A)$	$U_B(\alpha_A)$	$U_C(\alpha_A)$
$\{C, C, NP, A\}$	$-d_{CA}$	$\lambda_B\tau$	τ
$\{C, D, NP, NA\}$	$-L$	T	$\lambda_C(L)$
$\{C, D, NP, A\}$	$-L - d_{CA}$	$T + \lambda_B\tau$	$\tau + \lambda_C(L)$
$\{D, D, NP, NA\}$	$G - L$	$G - L$	$\lambda_C(-(G - L)$
$\{D, D, NP, A\}$	$G - L - d_{CA}$	$G - L + \lambda_B\tau$	$\tau + \lambda_C(-(G - L))$

Table A.1: Utility for each player during action sets which punish A (α_A)

From the table above, we have excluded punishments on A in which B engages in costly sanctioning of C, as it does not impact the strength of the punishment on A and B has no incentive to pay this cost, as well as action sets in which B cooperates with A while A defects, as this is a net gain for A over α_{peace} in the PD. In other words, we exclude any action set in which the payoff for A from one of the components of the game strictly dominates the payoff from that component that A would receive under α_{peace}.

We follow the same procedure for punishment action sets for B and C, with the caveat that we include punishment action sets in which C continues attacking even though this creates component game payoffs for the punished player that strictly dominate the payoff from that component during peace because we the ability of A and B to punish without the cooperation of C is substantively meaningful for our application.

Action Set	$U_A(\alpha_B)$	$U_B(\alpha_B)$	$U_C(\alpha_B)$
$\{D, C, NP, NA\}$	G	$-L$	$\lambda_C(-G)$
$\{D, C, NP, A\}$	$G - d_{CA}$	$-L + \lambda_B\tau$	$\tau + \lambda_C(-G)$
$\{D, D, NP, NA\}$	$G - L$	$G - L$	$\lambda_C(L - G)$
$\{D, D, NP, A\}$	$G - L - d_{CA}$	$G - L + \lambda_B\tau$	$\tau + \lambda_C(L - G)$

Table A.2: Utility for each player during action sets which punish B (α_B)

Action Set	$U_A(\alpha_C)$	$U_B(\alpha_C)$	$U_C(\alpha_C)$
$\{C, C, P, A\}$	$-d_{CA}$	$\lambda_B\tau - k$	$\tau - d_{BC}$
$\{C, C, P, NA\}$	0	$-k$	$-d_{BC}$
$\{D, C, NP, A\}$	$G - d_{CA}$	$-L + \lambda_B\tau$	$\tau + \lambda_C(-G)$
$\{D, C, NP, NA\}$	G	$-L$	$\lambda_C(-G)$
$\{D, C, P, A\}$	$G - d_{CA}$	$-L + \lambda_B\tau - k$	$\tau + \lambda_C(-G) - d_{BC}$
$\{D, C, P, NA\}$	G	$-L - k$	$\lambda_C(-G) - d_{BC}$

Table A.3: Utility for each player during action sets which punish C (α_C)

Table A.3 makes it clear that there are two ways for the other players to harm C, either through mutual cooperation and punishment, or through allowing A to defect while B cooperates. It is apparent in the BC component game that B can directly harm C through its sanction move, with a cost of k and a magnitude of d_{BC}. Alternately, A can defect while B cooperates, in the AB Prisoner's dilemma game component. In this case, C's preferences over the outcome of the AB component are used as a mechanism of punishment. Because C gains positive utility when A's payout is mutual cooperation, and negative utility when A's payout is more than mutual cooperation, A and B can jointly act such that A receives its maximal payout from the Prisoner's Dilemma, which causes maximal negative to C. The effectiveness of such a strategy is contingent on the particular value of $lambda_C$.

Appendix B

A Equilibria in the Responsibility Game

Here we examine the conditions that enable the peaceful equilibrium, σ_{peace}, which we define as an equilibrium in which all i players play α^i from the action set α_{peace} in all rounds. When this is a stable subgame perfect nash equilibrium, A is able to hold B responsible for the actions of C. Intuitively, the following must be true for this to occur:

- A must be able to credibly threaten to punish B severely enough that B prefers to exert control over C and maintain peace rather than allow C to continue attacking A
- B must be able to credibly threaten to punish C severely enough to make C prefer not to attack in the first place.

More simply, A has to be able to make B *want* to control C and B has to be *able* to control B.

A generic player Y has no incentive to deviate from σ_{peace} when its expected payoff from the peaceful equilibria is greater than its expected payoff from a breakdown in peaceful play.

$$U_Y(\sigma_{peace}) \geq U_{Y_0}(\alpha_{deviate}) + U_{Y_{1\ldots n}}(\alpha_Y) + U_{Y_{n+1\ldots\infty}}(\alpha_{peace}) \tag{B.1}$$

By definition $U_Y(\alpha_{peace}) = 0$ so Equation B.1 simplifies to:

$$0 \geq U_{Y_0}(\alpha_{deviate}) + U_{Y_{1\ldots n}}(\alpha_Y) \tag{B.2}$$

B.1 B defects first

Using Equation B.2, for player B we derive the conditions under which B has no incentive to deviate from σ_{peace}. These conditions are dependent on the punishment action set being used by other players. As noted in A.2.3 several punishment action sets on B can be used against B if B defects. We denote $U_B(\alpha_B) = X$ for generality, where X can be B's utility from any action in Table A.2.

Proposition: B has no incentive to deviate from σ_{peace} if any deviation would be met with n rounds of α_B by the other players and:

$$-G \geq (X - G)\delta - \delta^{n+1}X$$

Proof: We take the exact same approach as above.

$$0 \geq G + \frac{\delta(1-\delta^n)}{1-\delta}X \tag{B.3}$$

$$0 \geq G(1-\delta) + \delta X - \delta^{n+1}X \tag{B.4}$$

$$-G \geq (X-G)\delta - \delta^{n+1}X \tag{B.5}$$

In the case where $X = G - L$ we have the standard prisoner's dilemma, and as $n \to \infty$, $\delta \geq \frac{G}{L}$. In the case where $X = (G-L) + \lambda_B\tau$, the condition for σ_{peace} to be stable when $n \to \infty$

$$\delta \geq \frac{G}{L - \lambda_B\tau} \tag{B.6}$$

We compare this to the threshold δ threshold in a standard prisoners dilemma. Here δ must be higher in the three player game with C attacking to for B to have no incentive to deviate from the peaceful strategy than it would need to be to enable cooperation with only two players. The larger τ or λ_B — which together comprise the benefit B gets from C attacking A — the higher δ must be for cooperation to be attractive.

When the above condition is met there is no incentive for B to deviate from σ_{peace}

B.2 C Defects First

We are particularly interested in the case when C considers defecting first by attacking A in the first component of game.

Using Equation B.2 we derive the conditions under which C has no incentive to deviate from σ_{peace}. These conditions are dependent on the punishment action set

being used by other players. As above, several punishment action sets on C can be used against C if C defects.

It is worth noting that in contrast to a defection by B, when C defects its punishment begins in the same round rather than the next round because of the ordering of the components of the game. Because of this we must decompose the set of possible punishment action sets on C $\boldsymbol{\alpha_C}$, given in Table A.3 into two sets $\boldsymbol{\alpha_{C_{NA}}}$ and $\boldsymbol{\alpha_{C_A}}$ in which $\alpha_C^C = NA$ in the former and $\alpha_C^C = A$ in the latter. That is, $\boldsymbol{\alpha_{C_{NA}}}$ is the set of punishment action sets in which C does not attack and $\boldsymbol{\alpha_{C_A}}$ is the set of action sets in which C attacks. For generality we denote $U_C(\alpha_C)$ as X if the punishment action may be drawn from the full punishment action set $\boldsymbol{\alpha_C}$ and as either X_{NA} or X_A where the punishment action set may be drawn from $\boldsymbol{\alpha_{C_{NA}}}$ and $\boldsymbol{\alpha_{C_N}}$, respectively. These payoffs are given in Table A.3.

Proposition: C has no incentive to deviate from σ_{peace} if any deviation would be met with n rounds of α_C by the other players and:

$$\delta \geq \frac{\tau + X_{NA}}{\tau}$$

or

$$-\tau \geq X_{NA}$$

depending on the punishment action set being used against C.

Proof: Using the same approach as before it's easy to see:

$$0 \geq \tau + X_{NA} + \frac{\delta(1-\delta^n)}{1-\delta}X \tag{B.7}$$

This leads to two possibilities, either $X = X_{NA}$ or $X = X_A$. If $X = X_{NA}$ then:

$$0 \geq \tau + \frac{1-\delta^n}{1-\delta} X_{NA} \tag{B.8}$$

$$0 \geq \tau(1-\delta) + (1-\delta^n) X_{NA} \tag{B.9}$$

$$\delta \geq \frac{\tau + X_{NA} - \delta^n X_{NA}}{\tau} \tag{B.10}$$

Under the grim trigger strategy as $n \to \infty$ this reduces to:

$$\delta \geq \frac{\tau + X_{NA}}{\tau}$$

Because X_{NA} is always negative, this can be interpreted as indicating that as the magnitude of the punishment becomes larger, it requires a lower discount factor to maintain the ability to disincentivize C from attacking.

If $X = X_A$ then:

$$0 \geq \frac{1-\delta^n}{1-\delta} X_A \tag{B.11}$$

$$0 \geq \frac{\tau}{1-\delta} + \frac{1-\delta^n}{1-\delta} X_{NA} \tag{B.12}$$

$$0 \geq \tau + (1-\delta^n) X_{NA} \tag{B.13}$$

$$-\tau \geq X_{NA} - \delta^n X_{NA} \tag{B.14}$$

In this situation the discount factor doesn't impact the incentives. Under the grim trigger strategy as $n \to \infty$ this reduces to:

$$-\tau \geq X_{NA}.$$

When these conditions are met B has an effective punishment against C and it is possible for B to disincentivize C.

B.3 Punishment Credibility

The sections above give the conditions under which B prefers peace, and B is capable of making C prefer peace as well. However, these conditions rely variously on the punishment action sets played by the other players. In this section we detail the conditions under which these threats made by A and B are credible.

Which strategy is played during the punishment phase is a substantive issue, because which is used has implications for which conditions support peaceful Nash Equilibria, and sub-game perfection is a function of whether this punishment is itself supportable. Depending on which punishment strategy is played in the punishment phase, the punishment will be more or less severe on a defecting player. The stronger the punishment, the fewer rounds needed in the punishment phase to support the equilibria (if it is supportable at all).

In A we focused on a set of punishments, in which A and C are minimaxed and B is minimaxed by A but C continues attacking A regardless:

$$\alpha_{minimax_A} = \{D, D, NP, A\} \tag{B.15}$$

$$\alpha_{mimimax'_B} = \{D, D, NP, A\} \tag{B.16}$$

$$\alpha_{minimax_C} = \{D, C, P, A\} \tag{B.17}$$

The payoffs for these strategies, which we have discussed above, are reproduced here for convenience:

$$U_A(\alpha_{minimax_A}) = G - L - d_{CA} \tag{B.18}$$

$$U_B(\alpha_{minimax'_B}) = G - L + \lambda_B \tau \tag{B.19}$$

$$U_C \alpha_{minimax_C}) = \tau - d_{BC} + \lambda_C(-G) \tag{B.20}$$

By definition a strategy is a sub-game perfect nash equilibrium (SPNE) if the punishment threats used to sustain that equilibrium are credible. Therefore, following through with the punishment phase has no lower a expected payoff than not following though on punishment. These are credible when they are themselves SPNE strategies. Playing a strategy which is NE in the stage game in every round is necessarily SPNE. Both $\alpha_{minimax_A}$ and $\alpha_{mimimax'_B}$ are NE in the stage game, and thus are SPNE credible threats. It only remains to be seen if $\alpha_{minimax_C}$ is SPNE under any conditions. If not, we must turn to one of the other punishments enumerated in the preceding section and re-solve the game for the conditions which enable peace.

As $\alpha_{minimax_C}$ is a minimax strategy on C we know C has no incentive to change its choice; it is best responding to the minimization strategies of A and B. We examine the payoffs for A and B from playing $\alpha_{minimax_C}$:

$$U_A(\alpha_{minimax_C}) = G - d_{CA} \tag{B.21}$$

$$U_B(\alpha_{minimax_C}) = -L - k + \lambda_B \tau \tag{B.22}$$

A has no incentive to deviate here in the stage game when this strategy is being played as it is receiving the maximum possible payoff from the Prisoners' Dilemma, and its only move is in the Prisoners' Dilemma.

B has an incentive to deviate in the stage game. It could make itself better off by playing defect and/or by not attacking C. Thus, a strategy with this a punishment

phase where $n \to \infty$ (grim trigger) is necessarily not SPNE. We next check the minimal length punishment phase (n=1), if this strategy is not SPNE under these conditions then it is never SPNE.

$$\frac{(1-\delta^n)(-L) - k + \lambda_B \tau}{1-\delta} \geq \frac{G - L + \lambda_B \tau}{1-\delta} \tag{B.23}$$

Proposition: B has no incentive to deviate from playing the one round $\alpha_{minimax_C}$ punishment iff:

$$\delta \geq \frac{k+G}{k+L-\lambda_B \tau} \text{and} \tag{B.24}$$

$$\tau \leq \frac{L-G}{\lambda_B} \tag{B.25}$$

That is, when the B's payoff from a single punishment and the peaceful payoff in all future rounds exceeds B's payoff from not punishing and receiving the deviation payoff, from strategy $\alpha_{deviation} = \{D, DA, NP\}$ in all future rounds, then $\alpha_{minimax_C}$ is a credible threat. This reduces to a credibility condition in which the difference between mutual cooperation and defection must be greater than or equal to B's weighted gain from C attacking A.

Similarly, B has no incentive to deviate from playing the n round $\alpha_{minimax_C}$ punishment iff:

$$-L + \delta^n (k + L - \tau \lambda_B) \geq k + G - L \tag{B.26}$$

Proof of the $n = 1$ case:

$$-L - k + \lambda_B \tau \geq \frac{G - L + \lambda_B \tau}{1 - \delta} \tag{B.27}$$

$$(1 - \delta)(-L - k + \lambda_B \tau) \geq G - L + \lambda_B \tau \tag{B.28}$$

$$(1 - \delta)(-L - k) - \delta \lambda_B \tau \geq G - L \tag{B.29}$$

$$\delta(k + L - \lambda_B \tau) \geq k + G \tag{B.30}$$

and if $(k + L - \lambda_B \tau) > 0$ then

$$\delta \geq \frac{k + G}{k + L - \lambda_B \tau} \tag{B.31}$$

In the extreme, if δ was 1, and players value the future just as much as the past, then the threat would be credible when $\lambda_B \tau \leq L - G$.

Next we explore an alternate punishment action set from Table A.3, from above, in which: A cooperates, B cooperates, B sanctions C. We will call this single round punishment phase with this strategy, α_{pcl} (punish C lightly). The single round payoffs for this strategy are reproduced below for convenience.

$$U_A(\alpha_{pcl}) = -d_{CA} \tag{B.32}$$

$$U_B(\alpha_{pcl}) = -k + \lambda_B \tau \tag{B.33}$$

$$U_C(\alpha_{pcl}) = \tau - d_{BC} \tag{B.34}$$

When $d_{BC} \geq \tau$ this threat is effective at incentivizing C not to attack A. Now we

will turn to credibility. α_{pcl} requires no special action by A other than the strategy it was already playing during α_{peace}. Therefore, if α_{peace} was a NE, then α_{pcl} is credible as far as A is concerned. C is also already best responding, and has no incentive to move. Therefore, whether α_{pcl} is ultimately SPNE depends on whether B's threat on C is credible. For an n round punishment phase using this same punishment strategy, α_{pcl}, the threat is credible when the punishment strategy is preferable to the defection strategy for B, such that $U_B(\alpha_{pcl}) \geq U_B(\alpha_{defect})$:

$$\frac{(1-\delta^n)(-k+\lambda_B\tau)}{1-\delta} + 1 - \delta \geq \frac{G-L+\lambda_B\tau}{1-\delta} \tag{B.35}$$

Proposition: B has no incentive to deviate from playing the one round s_{pcl} punishment if and only if one or more of the four following conditions are met:

$$k < G-L \text{ and } \tau < \frac{G-L}{\lambda_B} \tag{B.36}$$

$$k < G-L \text{ and } \tau > \frac{G-L}{\lambda_B} \text{ and } \delta \geq \frac{k-(G-L)}{k-\lambda_B\tau} \tag{B.37}$$

$$k = G-L \text{ and } \tau \leq \frac{k}{\lambda_B} \tag{B.38}$$

$$k > G-L \text{ and } \tau < \frac{G-L}{\lambda_B} \text{ and } \delta \geq \frac{k-(G-L)}{k-\lambda_B\tau} \tag{B.39}$$

The third condition in B.38 can be written as a condition on k, $\frac{\delta\lambda_B\tau+G-L}{1-\delta} \geq k$ if $k-\lambda_B\tau > 0$, otherwise $\frac{\delta\lambda_B\tau+G-L}{1-\delta} \leq k$.

Proof: Each possible outcome was checked exhaustively using symbolic algebraic software Mathematica [295]. The code can be found at `http://cs.unm.edu/~bedwards/data/EquilibriaRespGame.nb`.

In the general case of n, B has no incentive to deviate from α_{pcl} during the punishment phase for the same conditions with the exception that:

$$\delta^n \geq \frac{k-(G-L)}{k-\lambda_B \tau} \tag{B.40}$$

for equations B.37 and B.39.

During the single round punishment phase, B must only pay for k to punish in one round, but this is balanced against the gains to B from $\lambda_B \tau$ during defection. Thus, the larger either λ_B or τ are, the larger the difference $L-G$ must be for the threat to be credible in a one round phase, and the peaceful equilibrium it supported is SPNE.

As $n \to \infty$, which is a grim trigger in the limit, α_{pcl}, is credible under the following condition:

$$G - L \geq k \tag{B.41}$$

Punishment strategies $\alpha_{minimax_A}$, $\alpha_{minimax'_B}$, and α_{pcl} with an n round punishment phase are credible punishments which will support a peaceful equilibrium when $n \to \infty$ (grim trigger) and $G - L \geq k$. The intuition here is that this punishment is credible when the benefit of mutual cooperating over mutual defection is larger than the cost of punishing C. Other length punishments follow a similar intuition, but deal with B gaining some benefit from C's attack on A, $\lambda_B \tau$ during defection and punishment rounds weighed against future peaceful play following the punishment round.

Credibility conditions can be computed similarly for each of the punishment action sets detailed in section 3.

B.4 A Holding B Responsible for the Actions of C

There are a wide variety of strategies that can be used to support σ_{peace}, and the particular parametric conditions are dependent upon which punishment action sets are being used during punishment phases. These conditions are computed in the manner detailed above. A must have a punishment action set on B that is both effective and credible, and A and B must have a punishment action set on C that is both effective and credible. When all of these are true, C is disincentivized from ever defecting in the first place.

References

[1] 164.400-414, . C. S. Health insurance portability and accountability act, Aug. 1996.

[2] Acquisti, A., Friedman, A., and Telang, R. Is there a cost to privacy breaches? an event study. *ICIS 2006 Proceedings* (2006), 94.

[3] Adamic, L. A., and Huberman, B. A. Power-law distribution of the world wide web. *Science 287* (2000), 2115.

[4] Alkabani, Y., and Koushanfar, F. N-variant ic design: methodology and applications. In *Proceedings of the 45th annual Design Automation Conference* (2008), ACM, pp. 546–551.

[5] Anderson, R., Barton, C., Böhme, R., Clayton, R., van Eeten, M., Levi, M., Moore, T., and Savage, S. Measuring the cost of cybercrime. In *WEIS* (2012).

[6] Ardagna, C. A., Cremonini, M., Damiani, E., Di Vimercati, S. D. C., and Samarati, P. Location privacy protection through obfuscation-based techniques. In *Data and Applications Security XXI*. Springer, 2007, pp. 47–60.

[7] Arnold, T. B., and Emerson, J. W. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal 3*, 2 (2011), 34–39.

[8] Arora, A., Krishnan, R., Telang, R., and Yang, Y. An empirical analysis of vendor response to disclosure policy. In *WEIS* (2005).

[9] Asghari, H., Ciere, M., and Van Eeten, M. J. Post-mortem of a zombie: conficker cleanup after six years. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 1–16.

[10] Asteriou, D., and Hall, S. G. *Applied Econometrics: a modern approach using eviews and microfit*. Palgrave Macmillan New York, 2007.

[11] Aviv, A. J., Budzitowski, D., and Kuber, R. Is bigger better? comparing user-generated passwords on 3x3 vs. 4x4 grid sizes for android's pattern unlock. In *Proceedings of the 31st Annual Computer Security Applications Conference* (2015), ACM, pp. 301–310.

[12] Axelrod, R., and Hamilton, W. D. The evolution of cooperation. *Science 211*, 4489 (1981), 1390–1396.

[13] Backes, M., Bugiel, S., Hammer, C., Schranz, O., and von Styp-Rekowsky, P. Boxify: Full-fledged app sandboxing for stock android. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 691–706.

[14] Bagchi, K., and Udo, G. An analysis of the growth of computer and internet security breaches. *Communications of the Association for Information Systems 12*, 1 (2003), 46.

[15] Bagci, I. E., Roedig, U., Martinovic, I., Schulz, M., and Hollick, M. Using channel state information for tamper detection in the internet of things. In *Proceedings of the 31st Annual Computer Security Applications Conference* (2015), ACM, pp. 131–140.

[16] Bank, W. World bank data. http://data.worldbank.org/, Mar. 2015.

[17] Barrantes, E. G., Ackley, D. H., Palmer, T. S., Stefanovic, D., and Zovi, D. D. Randomized instruction set emulation to disrupt binary code injection attacks. In *Proceedings of the 10th ACM conference on Computer and communications security* (2003), ACM, pp. 281–289.

[18] Barrett, D., and Gorman, S. U.s. charges five in chinese army with hacking. *Wall Street Journal* (May 2014).

[19] Barrett, D., Yadron, D., and Paletta, D. U.s. suspects hackers in china breached about 4 million people's records, officials say. *Wall Street Journal* (June 2015).

[20] Barth, A., Jackson, C., Reis, C., and Team, T. The security architecture of the chromium browser, 2008.

[21] Barthe, G., Dupressoir, F., Grégoire, B., Kunz, C., Schmidt, B., and Strub, P.-Y. Easycrypt: A tutorial. In *Foundations of Security Analysis and Design VII*. Springer, 2014, pp. 146–166.

[22] Bayes, C. L., and Branco, M. Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics 21*, 2 (2007), 141–163.

[23] Bérard, B., Bidoit, M., Finkel, A., Laroussinie, F., Petit, A., Petrucci, L., and Schnoebelen, P. *Systems and software verification: model-checking techniques and tools*. Springer Science & Business Media, 2013.

[24] Bergsma, F., Dowling, B., Kohlar, F., Schwenk, J., and Stebila, D. Multi-ciphersuite security of the secure shell (ssh) protocol. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014), ACM, pp. 369–381.

[25] Beringer, L., Petcher, A., Katherine, Q. Y., and Appel, A. W. Verified correctness and security of openssl hmac. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 207–221.

[26] Bishop, C. M., et al. *Pattern recognition and machine learning*, vol. 1. springer New York, 2006.

[27] Bishop, M. A. *The Art and Science of Computer Security*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

[28] Blakley, B., McDermott, E., and Geer, D. Information security is information risk management. In *Proceedings of the 2001 workshop on New security paradigms* (2001), ACM, pp. 97–104.

[29] Bosworth, M. H. Tjx data breach victims reach 94 million. *Consumer Affairs* (Oct. 2007).

[30] BotFrei. botfrei.de: The anti-botnet advisory centre. `https://www.botfrei.de/`, May 2014.

[31] Boyer, R. S., and Moore, J. S. Proof checking the rsa public key encryption algorithm. *The American Mathematical Monthly 91*, 3 (1984), 181–189.

[32] Brockwell, P. J., and Davis, R. A. *Time series: theory and methods*. Springer, 2009.

[33] Burr, S. R. Cybersecurity information sharing act, Oct. 2015.

[34] Cao, Y., Zhang, H., Zhao, X., and Yu, H. Video steganography based on optimized motion estimation perturbation. In *Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security* (2015), ACM, pp. 25–31.

[35] Cebula, J., and Young, L. A taxonomy of operational cyber security risks. Tech. rep., DTIC Document, 2010.

[36] Center, M. I. Apt1: Exposing one of chinas cyber espionage units. *Mandian. com* (2013).

[37] Chang, H., Jamin, S., Mao, Z. M., and Willinger, W. An empirical approach to modeling inter-as traffic matrices. In *Proc. of ACM IMC* (2005), USENIX Association.

[38] Chen, B. X. Home depot investigates a possible credit card breach. *The New York Times* (Sept. 2014).

[39] Chen, L., and Avizienis, A. N-version programming: A fault-tolerance approach to reliability of software operation. In *Digest of Papers FTCS-8: Eighth Annual International Conference on Fault Tolerant Computing* (1978), pp. 3–9.

[40] Chen, Y.-F., Hsu, C.-H., Lin, H.-H., Schwabe, P., Tsai, M.-H., Wang, B.-Y., Yang, B.-Y., and Yang, S.-Y. Verifying curve25519 software. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014), ACM, pp. 299–309.

[41] Chen, Z., Ji, C., and Barford, P. Spatial-temporal characteristics of internet malicious sources. In *INFOCOM* (2008), IEEE.

[42] Cho, C. Y., Shin, E. C. R., Song, D., et al. Inference and analysis of formal models of botnet command and control protocols. In *Proc. of ACM CCS* (2010), ACM.

[43] Chrobok, N., Trotman, A., and O'Keefe, R. Advantages and vulnerabilities of pull-based email-delivery. In *Proceedings of the Eighth Australasian Conference on Information Security-Volume 105* (2010), Australian Computer Society, Inc., pp. 22–31.

[44] Claburn, T. Most security breaches go unreported. *Information Week* (July 2008).

[45] Clark, D. D., and Landau, S. Untangling attribution. *Harv. Nat'l Sec. J. 2* (2011), 323.

[46] Clark, S., Collis, M., Blaze, M., and Smith, J. M. Moving targets: Security and rapid-release in firefox. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014), ACM, pp. 1256–1266.

[47] Clarke, R. A., and Knake, R. *Cyber War: The Next Threat to National Security and What to Do About It.* Ecco, 2010.

[48] Clauset, A., Shalizi, C., and Newman, M. Power-law distributions in empirical data. *Arxiv preprint arxiv:0706.1062* (2007).

[49] Clauset, A., Woodard, R., et al. Estimating the historical and future probabilities of large terrorist events. *The Annals of Applied Statistics 7*, 4 (2013), 1838–1865.

[50] Clearinghouse, P. R. Mission statement. `https://www.privacyrights.org/content/about-privacy-rights-clearinghouse#goals`, May 2014.

[51] Clearinghouse, P. R. Chronology of data breaches: Faq. `https://www.privacyrights.org/content/chronology-data-breaches-faq`, 2015.

[52] Collins, M. P., Shimeall, T. J., Faber, S., Janies, J., Weaver, R., De Shon, M., and Kadane, J. Using uncleanliness to predict future botnet addresses. In *Proc. of ACM IMC* (2007).

[53] Condon, E., He, A., and Cukier, M. Analysis of computer security incident data using time series models. In *Software Reliability Engineering, 2008. ISSRE 2008. 19th International Symposium on* (2008), IEEE, pp. 77–86.

[54] Cooper, R. *Coordination games*. Cambridge University Press, 1999.

[55] Corporation, S. Internet security threat report 2014. `http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf`, Apr. 2014.

[56] Corrons, L. Mariposa botnet. *Panda Labs* (Mar. 2010).

[57] Costa, M., Crowcroft, J., Castro, M., Rowstron, A., Zhou, L., Zhang, L., and Barham, P. Vigilante: End-to-end containment of internet worms. In *ACM SIGOPS Operating Systems Review* (2005), vol. 39, ACM, pp. 133–147.

[58] Cova, M., Kruegel, C., and Vigna, G. Detection and analysis of drive-by-download attacks and malicious JavaScript code. In *Proc. WWW '10* (2010), pp. 281–290.

[59] Covington, and LLP, B. Data breach notification bills introduced in house and senate. *The National Law Review* (Feb. 2015).

[60] Cox, B., Evans, D., Filipi, A., Rowanhill, J., Hu, W., Davidson, J., Knight, J., Nguyen-Tuong, A., and Hiser, J. N-variant systems: a secretless framework for security through diversity. In *Usenix Security* (2006), vol. 6, pp. 105–120.

[61] Cranor, L. F., and LaMacchia, B. A. Spam! *Communications of the ACM 41*, 8 (1998), 74–83.

[62] Creswell, J., and Dash, E. Banks unsure which cards were exposed in breach. *The New York Times* (June 2005).

[63] Curtin, M., and Ayres, L. T. Using science to combat data loss: Analyzing breaches by type and industry. *ISJLP 4* (2008), 569.

[64] Curtsinger, C., Livshits, B., Zorgn, B., and Seifert, C. ZOZZLE: Fast and precise in-browser JavaScript malware detection. In *Proc. 20th USENIX Security Symp.* (Aug. 2011).

[65] d'Agapeyeff, A. *Codes and Ciphers-A History Of Cryptography*. Hesperides Press, 2008.

[66] Dagon, D., Zou, C. C., and Lee, W. Modeling botnet propagation using time zones. In *NDSS* (2006), vol. 6.

[67] Davis, J. H. Hacking of government computers exposed 21.5 million people. *The New York Times* (July 2015).

[68] Demme, J., Maycock, M., Schmitz, J., Tang, A., Waksman, A., Sethumadhavan, S., and Stolfo, S. On the feasibility of online malware detection with performance counters. *ACM SIGARCH Computer Architecture News 41*, 3 (2013), 559–570.

[69] Dhamdhere, A., and Dovrolis, C. Ten years in the evolution of the internet ecosystem. In *Proc. of ACM IMC* (2008), ACM.

[70] Dickey, D. A., and Fuller, W. A. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: Journal of the Econometric Society* (1981).

[71] Director of National Intelligence James R. Clapper. Statement by director of national intelligence james r. clapper on allegations of economic espionage. `http://www.dni.gov/index.php/newsroom/press-releases/191-press-releases-2013/926-statement-by-director-of-national-intelligence-james-r-clapper-on-allegations-of-economic-espionage`, Sept. 2013.

[72] Donohue, B. Kapersky knocks down kelihos botnet again, but expects return. *Threatpost.com* (Mar. 2012).

[73] Edwards, B., and Forrest, S. How do complex systems protect themselves from malicious behavior. In *Conference on Complex Systems (to appear)* (Sept. 2015).

[74] Edwards, B., Hofmeyr, S., and Forrest, S. Hype and heavy tails: A closer look at data breaches. In *WEIS* (2015).

[75] Edwards, B., Hofmeyr, S., Forrest, S., and van Eeten, M. Analyzing and modeling longitudinal security data: Promise and pitfalls. In *Proceedings of the 31st Annual Computer Security Applications Conference* (2015), ACM, pp. 391–400.

[76] Edwards, B., Moore, T., Stelle, G., Hofmeyr, S., and Forrest, S. Beyond the blacklist: modeling malware spread and the effect of interventions. In *Proceedings of the 2012 workshop on New security paradigms* (2012), ACM, pp. 53–66.

[77] Edwards, B., Moore, T., Stelle, G., Hofmeyr, S., and Forrest, S. Beyond the blacklist: modeling malware spread and the effect of interventions. In *Proceedings of the 2012 workshop on New security paradigms* (New York, NY, USA, 2012), NSPW '12, ACM, pp. 53–66.

[78] Edwards, B., Moore, T., Stelle, G., Hofmeyr, S., and Forrest, S. Beyond the blacklist: Modeling malware spread and the effects of interventions. In *Proc. of NSPW 2012* (2012).

[79] Egelman, S., Cranor, L. F., and Hong, J. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2008), CHI '08, ACM, pp. 1065–1074.

[80] Ellison, G. Cooperation in the prisoner's dilemma with anonymous random matching. *The Review of Economic Studies 61*, 3 (1994), 567–588.

[81] Emerson, E. A., and Clarke, E. M. Using branching time temporal logic to synthesize synchronization skeletons. *Science of Computer programming 2*, 3 (1982), 241–266.

[82] Erasmus, J. Compromised ftp details being exploited by in the wild malware, June 2009. `http://www.prevx.com/blog/132/Compromised.html`.

[83] Espiner, T. Dutch police take down bredolab botnet. *ZDNet* (Oct. 2010).

[84] Esponda, F., Forrest, S., and Helman, P. Negative representations of information. *International Journal of Information Security 8*, 5 (2009), 331–345.

[85] Falliere, N., Murchu, L. O., and Chien, E. W32. stuxnet dossier. *White paper, Symantec Corp., Security Response* (2011).

[86] Farwell, J. P., and Rohozinski, R. Stuxnet and the future of cyber war. *Survival 53*, 1 (2011), 23–40.

[87] Fayaz, S. K., Tobioka, Y., Sekar, V., and Bailey, M. Bohatei: flexible and elastic ddos defense. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 817–832.

[88] Fearon, J. D. Signaling foreign policy interests tying hands versus sinking costs. *Journal of Conflict Resolution 41*, 1 (1997), 68–90.

[89] Fearon, J. D., and Laitin, D. D. Explaining interethnic cooperation. *American political science review 90*, 04 (1996), 715–735.

[90] Finkle, J. Experian enmeshed in litigation over business that was breached. *Reuters* (Apr. 2014).

[91] Friedman, J. W. A non-cooperative equilibrium for supergames. *The Review of Economic Studies* (1971), 1–12.

[92] Geer, D. Monoculture on the back of the envelope. *Login 30*, 6 (2005), 6–8.

[93] Geer, D. Cybersecurity as realpolitik. `http://geer.tinho.net/geer.blackhat.6viii14.txt`, Aug. 2014.

[94] Geer, D., and Charney, A. R. Debate: Is an operating system monoculture a threat to security? *Proceedings of the USENIX Annual Technical Conference* (June 2004).

[95] Gelman, A., and Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[96] Gemalto. 2015 first half review. *Findings from the Breach Level Index* (Sept. 2015).

[97] Georgiev, M., Iyengar, S., Jana, S., Anubhai, R., Boneh, D., and Shmatikov, V. The most dangerous code in the world: validating ssl certificates in non-browser software. In *Proceedings of the 2012 ACM conference on Computer and communications security* (New York, NY, USA, 2012), CCS '12, ACM, pp. 38–49.

[98] Ghosh, A., and Schwartzbard, A. A study in using neural networks for anomaly and misuse detection. In *Proc. USENIX Security Symp.* (1999).

[99] Gnedenko, B., Kolmogorov, A., Chung, K., and Doob, J. *Limit distributions for sums of independent random variables*, vol. 195. Addison-Wesley Reading, MA:, 1968.

[100] Goldberg, S., and Forrest, S. Implications of security enhancements and interventions for core internet infrastructure. In *Telecommunications Policy Research Conference (TPRC'42), Arlington, VA* (2014).

[101] Good, D. I. The foundations of computer security: We need some. `http://www.ieee-security.org/CSFWweb/goodessay.html`, Sept. 1986.

[102] Goodin, D. Waledac botnet 'decimated' by ms takedown. *The Register* (Mar. 2010).

[103] Goodin, D. "slain" kelihos botnet still spams from beyond the grave. *arstechnica* (Feb. 2012).

[104] Google. Safe browsing api. `http://code.google.com/apis/safebrowsing/`.

[105] Google. An update on our war against account hijackers. *Google Official Blog* (Feb. 2013).

[106] Gorodetski, V., and Kotenko, I. Attacks against computer network: Formal grammar-based framework and simulation tool. In *Recent Advances in Intrusion Detection*, A. Wespi, G. Vigna, and L. Deri, Eds., vol. 2516 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2002, pp. 219–238.

[107] Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969).

[108] Graves, J. T., Acquisti, A., and Christin, N. Should payment card issuers reissue cards in response to a data breach? *2014 Workshop on the Economics of Information Security* (2014).

[109] Gu, G., Zhang, J., and Lee, W. Botsniffer: Detecting botnet command and control channels in network traffic. In *Proceedings of the 15th Annual Network and Distributed System Security Symposium (NDSS08)* (2008), San Diego, CA.

[110] Gudkova, D. Kaspersky security bulletin: Spam evolution 2013. `https://securelist.com/analysis/kaspersky-security-bulletin/36843/kaspersky-security-bulletin-spam-evolution-2012/`, Jan. 2013.

[111] Haight, F. A. *Handbook of the Poisson distribution*. Wiley New York, 1967.

[112] Hannas, W. C., Mulvenon, J., and Puglisi, A. B. *Chinese industrial espionage: Technology acquisition and military modernisation*. Routledge, 2013.

[113] Hansman, S., and Hunt, R. A taxonomy of network and computer attacks. *Computers & Security 24*, 1 (2005), 31 – 43.

[114] Hcine, M. B., and Bouallegue, R. Fitting the log skew normal to the sum of independent lognormals distribution. *arXiv preprint arXiv:1501.02344* (2015).

[115] Hess, A., Jung, M., and Schfer, G. Fidran: A flexible intrusion detection and response framework for active networks. In *ISCC '03* (2003).

[116] Hird, S. Technical solutions for controlling spam. *proceedings of AUUG2002* (2002).

[117] Hofmeyr, S., Moore, T., Edwards, B., Forrest, S., and Stelle, G. Modeling Internet-scale policies for cleaning up malware. In *Proc. 10th Workshop on the Economics of Information Security* (2011).

[118] Hofmeyr, S., Moore, T., Forrest, S., Edwards, B., and Stelle, G. Modeling internet-scale policies for cleaning up malware. In *Economics of Information Security and Privacy III*. Springer, 2013.

[119] Hofmeyr, S. A., Moore, T., Forrest, S., Edwards, B., and Stelle, G. Modeling internet-scale policies for cleaning up malware. In *WEIS* (2011).

[120] Homan, M. D., and Gelman, A. The No-U-Turn Sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research 15*, 1 (2014), 1593–1623.

[121] Horowitz, M. C. The rise and spread of suicide bombing. *Annual Review of Political Science 18* (2015), 69–84.

[122] Howard, J., Longstaff, T., et al. A common language for computer security incidents. *Sandia Report: SAND98-8667, Sandia National Laboratories, http://www. cert. org/research/taxonomy_988667. pdf* (1998).

[123] HP. Immunity manager. Website. http://www.hp.com/rnd/pdfs/ProCurve_Network_Immunity_Manager1_0.pdf.

[124] Huq, N. Follow the data: Dissecting data breaches and debunking myths. *TrendMicro Research Paper* (Sept. 2015).

[125] Igure, V., and Williams, R. Taxonomies of attacks and vulnerabilities in computer systems. *Communications Surveys Tutorials, IEEE 10*, 1 (quarter 2008), 6 –19.

[126] Inc., R. Redspin breach report 2011: Protected health information. `http://www.redspin.com/docs/Redspin_PHI_2011_Breach_Report.pdf`, Feb. 2012.

[127] Inc., R. Redspin breach report 2012: Protected health information. `http://www.redspin.com/docs/Redspin_Breach_Report_2012.pdf`, Feb. 2013.

[128] Inc., R. Redspin breach report 2013: Protected health information. `https://www.redspin.com/docs/Redspin-2013-Breach-Report-Protected-Health-Information-PHI.pdf`, Feb. 2014.

[129] Ioannidis, J., and Bellovin, S. M. Implementing pushback: Router-based defense against ddos attacks, 2002.

[130] Jacobs, J. Analyzing ponemon cost of data breach. `http://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/`, Dec. 2014.

[131] John, J., Yu, F., Xie, Y., Abadi, M., and Krishnamurthy, A. deSEO: Combating search-result poisoning. In *Proceedings of the USENIX Security Symposium 2011* (San Francisco, CA, 2011).

[132] Johnson, B., Chuang, J., Grossklags, J., and Christin, N. Metrics for measuring isp badness: The case of spam. In *Financial Cryptography and Data Security*. Springer, 2012.

[133] Kalakota, P., and Huang, C.-T. On the benefits of early filtering of botnet unwanted traffic. In *Proc. of ICCCN* (2009).

[134] Kaloper-Meršinjak, D., Mehnert, H., Madhavapeddy, A., and Sewell, P. Not-quite-so-broken tls: lessons in re-engineering a security protocol specification and implementation. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 223–238.

[135] Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., and Savage, S. Spamalytics: An empirical analysis of spam marketing conversion. In *Proc. ACM CCS* (2008), ACM.

[136] Kanich, C., Weaver, N., McCoy, D., Halvorson, T., Kreibich, C., Levchenko, K., Paxson, V., Voelker, G. M., and Savage, S. Show me the money: Characterizing spam-advertised revenue. In *USENIX* (2011).

[137] Karasaridis, A., Rexroad, B., and Hoeflin, D. Wide-scale botnet detection and characterization. In *Proc. of HotBots* (2007), Cambridge, MA.

[138] Karlin, J., Rexford, J., and Forrest, S. Pretty good bgp: Improving bgp by cautiously adopting routes. In *Proc. CNP '06* (2006).

[139] Keizer, G. Rustock take-down proves botnets can be crippled, says microsoft. *Computer World* (July 2011).

[140] KENDALL, M., et al. Rank correlation methods. *Rank correlation methods.* (1948).

[141] Khan, M. T., Huo, X., Li, Z., and Kanich, C. Every second counts: Quantifying the negative externalities of cybercrime via typosquatting. In *Security and Privacy (SP), 2015 IEEE Symposium on* (2015), IEEE, pp. 135–150.

[142] Killourhy, K., Maxion, R., and Tan, K. A defense-centric taxonomy based on attack manifestations. In *Dependable Systems and Networks, 2004 International Conference on* (june-1 july 2004), pp. 102 – 111.

[143] Kohavi, R., et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI* (1995), vol. 14.

[144] Kokkodis, M., and Faloutsos, M. Spamming botnets: Are we losing the war. *Proceedings of CEAS 2009* (2009).

[145] Kosseff, J. Analysis of White House data breach notification bill. *The National Law Review* (Jan. 2015).

[146] Kotenko, I., Konovalov, A., and Shorov, A. Agent-based modeling and simulation of botnets and botnet defense. In *Conf. on Cyber Conflict* (2010).

[147] Krashakov, S. A., Teslyuk, A. B., and Shchur, L. N. On the universality of rank distributions of website popularity. *Computer Networks 50* (August 2006), 1769–1780.

[148] Krebs, B. Host of internet spam groups is cut off. *The Washington Post* (Nov. 2008).

[149] Krebs, B. Organized crime behind a majority of data breaches. *The Washington Post* (Apr. 2009).

[150] Krebs, B. Payment processor breach may be largest ever. *The Washington Post* (Jan. 2009).

[151] Krebs, B. The scrap value of a hacked pc. *The Washington Post* (May 2009).

[152] Krebs, B. Microsoft ambushes waledac botnet, shutters whistleblower site. *Krebs on Security* (Feb. 2010).

[153] Krebs, B. Takedowns: The shuns and stuns that take the fight to the enemy. *McAfee Security Journal 6* (2010).

[154] Krebs, B. U.s. government takes down coreflood botnet. *Krebs on Security* (2011).

[155] Krebs, B. Rogue pharma, fake av vendors feel credit card crunch. *Krebs On Security* (Oct. 2012).

[156] Krebs, B. Oracle ships critical security update for java. *Krebs on Security* (Jan. 2013).

[157] Krebs, B. Polish takedown targets 'virut' botnet. *Krebs On Security* (Jan. 2013).

[158] Krebs, B. What you need to know about the java exploit. *Krebs on Security* (Jan. 2013).

[159] Krebs, B. Home depot: Hackers stole 53m email addresses. *Krebs on Security* (Nov. 2014).

[160] Kropf, T. *Introduction to formal hardware verification*. Springer Science & Business Media, 2013.

[161] Kupsch, J. A., and Miller, B. P. Why do software assurance tools have problems finding bugs like heartbleed? *Continuous Software Assurance Marketplace 22* (2014).

[162] Kwon, J., and Johnson, M. E. The market effect of healthcare security: Do patients care about data breaches? In *Workshop on the Economics of Information Security (WEIS 15)* (2015).

[163] Laube, S., and Böhme, R. The economics of mandatory security breach reporting to authorities. In *Workshop on the Economics of Information Security (WEIS), Delft* (2015).

[164] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (2005), ACM, pp. 49–60.

[165] Lekies, S., Stock, B., Wentzel, M., and Johns, M. The unexpected dangers of dynamic javascript. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 723–735.

[166] Leontiadis, N., Moore, T., and Christin, N. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014), ACM, pp. 930–941.

[167] Levchenko, K., Chachra, N., Enright, B., Felegyhazi, M., Grier, C., Halvorson, T., Kanich, C., Kreibich, C., Liu, H., McCoy, D., Pitsillidis, A., Weaver, N., Paxson, V., Voelker, G., and Savage, S. Click trajectories: End-to-end analysis of the spam value chain. In *Proc. IEEE Sym. and Security and Privacy* (Oakland, CA, 2011).

[168] Leydon, J. Google: Botnet takedowns fail to stem spam tide. `http://www.theregister.co.uk/2010/04/18/google_botnet_takedowns/`, Apr. 2010.

[169] Li, Z., Liao, Q., and Striegel, A. Botnet economics: uncertainty matters. In *Managing Information Risk and the Economics of Security*. Springer, 2009.

[170] Lin, E., Greenberg, S., Trotter, E., Ma, D., and Aycock, J. Does domain highlighting help people identify phishing sites? In *Proceedings of the 2011 annual conference on Human factors in computing systems* (2011), CHI '11, pp. 2075–2084.

[171] Lindberg, G. Anti-spam recommendations for smtp mtas, 1999.

[172] Lindsay, J. R. Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack. *Journal of Cybersecurity* (2015).

[173] Liu, H., Levchenko, K., Félegyházi, M., Kreibich, C., Maier, G., Voelker, G. M., and Savage, S. On the effects of registrar-level intervention. In *Proc. USENIX Workshop on Large-scale Exploits and Emergent Threats (LEET)* (Boston, MA, March 2011).

[174] Liu, X., Yang, X., and Lu, Y. To filter or to authorize: Network-layer dos defense against multimillion-node botnets. In *ACM SIGCOMM* (2008), vol. 38, ACM.

[175] Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., and Liu, M. Cloudy with a chance of breach: Forecasting cyber security incidents. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 1009–1024.

[176] Liu, Y., Zhang, J., Sarabi, A., Liu, M., Karir, M., and Bailey, M. Predicting cyber security incidents using feature-based characterization of network-level malicious activities. In *Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics* (2015), ACM, pp. 3–9.

[177] LLC, M. Maxmind geoip, 2008.

[178] LLC, P. I. 2014 cost of data breach study: Global analysis. `http://www.ponemon.org/blog/ponemon-institute-releases-2014-cost-of-data-breach-global-analysis`, May 2014.

[179] Lu, L., Yegneswaran, V., Porras, P., and Lee, W. Blade: An attack-agnostic approach for preventing drive-by malware infection. *Proceedings of the 17th ACM Conference on Computer and Communications Security* (2010).

[180] Mahajan, R., Wetherall, D., and Anderson, T. Understanding bgp misconfiguration. In *ACM SIGCOMM* (2002), vol. 32, ACM.

[181] Maillart, T., and Sornette, D. Heavy-tailed distribution of cyber-risks. *The European Physical Journal B-Condensed Matter and Complex Systems 75*, 3 (2010), 357–364.

[182] Massey Jr, F. J. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association 46*, 253 (1951), 68–78.

[183] Mathews, A. W., and Yadron, D. Health insurer anthem hit by hackers. *The Wall Street Journal* (Feb. 2015).

[184] Maynor, D. *Metasploit Toolkit for Penetration Testing, Exploit Development, and Vulnerability Research*. Syngress, 2007.

[185] The Coq development team. *The Coq proof assistant reference manual*. LogiCal Project, 2004. Version 8.0.

[186] McCoy, D., Dharmdasani, H., Kreibich, C., Voelker, G. M., and Savage, S. Priceless: The role of payments in abuse-advertised goods. In *Proc. of ACM CCS* (2012).

[187] Meiss, M., Gonalves, B., Ramasco, J., Flammini, A., and Menczer, F. Modeling traffic on the web graph. In *Algorithms and Models for the Web-Graph*, R. Kumar and D. Sivakumar, Eds., vol. 6516 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2010, pp. 50–61.

[188] Microsoft. Microsoft academic search. `http://academic.research.microsoft.com/`.

[189] Microsoft. Microsoft security intelligence report. `http://download.microsoft.com/download/0/3/3/0331766E-3FC4-44E5-B1CA-2BDEB58211B8/Microsoft_Security_Intelligence_Report_volume_11_English.pdf`, Aug. 2011.

[190] Mirkovic, J., and Reiher, P. A taxonomy of ddos attack and ddos defense mechanisms. *SIGCOMM Comput. Commun. Rev. 34*, 2 (Apr. 2004), 39–53.

[191] Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet mathematics 1*, 2 (2004), 226–251.

[192] Moore, T., and Clayton, R. Discovering phishing dropboxes using email metadata. In *eCrime Researchers Summit (eCrime), 2012* (2012), IEEE.

[193] Moore, T., Leontiadis, N., and Christin, N. Fashion crimes: Trending-term exploitation on the web. In *Proc. ACM CCS'11* (Chicago, IL, Oct. 2011).

[194] Morrison, T. Spam botnets: The fall of grum and the rise of festi. *SpamHaus Blog* (Aug. 2012).

[195] Moshchuk, A., Bragin, T., Gribble, S. D., and Levy, H. M. A crawler-based study of spyware in the web. In *NDSS* (2006).

[196] Moura, G. C. *Internet bad neighborhoods*. No. 12 in University of Twente Dissertation. Giovane Cesar Moreira Moura, 2013.

[197] Murphy, K. Verisign demands website takedown powers. *The Register* (Oct. 2011).

[198] Nadji, Y., Antonakakis, M., Perdisci, R., Dagon, D., and Lee, W. Beheading hydras: performing effective botnet takedowns. In *Proc. of SIGSAC* (2013), ACM.

[199] Nakashima, E. Chinese breach of 4 million federal workers. *The Washington Post* (June 2015).

[200] Nakashima, E. Chinese government has arrested hackers it says breached opm database. *The Washington Post* (Dec. 2015).

[201] Nakashima, E. Following u.s. indictments, china shifts commercial hacking away from military to civilian agency. *The Washington Post* (Nov. 2015).

[202] Namestnikov, Y. The economics of botnets. *Analysis on Viruslist. com, Kapersky Lab* (2009).

[203] Narayanan, A., and Shmatikov, V. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105* (2006).

[204] Nations, U. Un geographic division. `http://millenniumindicators.un.org/unsd/methods/m49/m49regin.htm`, Oct. 2013.

[205] Neti, S., Somayaji, A., and Locasto, M. E. Software diversity: Security, entropy and game theory. In *Proceedings of the 7th USENIX Workshop on Hot Topics in Security, HotSec* (2012), vol. 12.

[206] Neti, S., Somayaji, A., and Locasto, M. E. Software diversity: Security, entropy and game theory. In *HotSec* (2012).

[207] News, S. F. A young botnet suspect arrested by russian authorities. *SpamFighter* (July 2012).

[208] Niels Provos. Safe browsing - protecting web users for 5 years and counting. `http://googleonlinesecurity.blogspot.com/2012/06/safe-browsing-protecting-web-users-for.html`.

[209] of Health, U. D., and Services, H. Annual report to congress on breaches of unsecured protected health information. `http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/breachreport2011-2012.pdf`, 2014.

[210] of Justice, U. D. Alleged international hacker indicted for massive attack on u.s. retail and banking networks. `http://www.justice.gov/opa/pr/alleged-international-hacker-indicted-massive-attack-us-retail-and-banking-networks`, Aug. 2009.

[211] of Justice, U. D. U.s. charges five chinese military hackers with cyber espionage against u.s. corporations and a labor organization for commercial advantage, May 2014.

[212] of Standards, N. I., and Technology. Framework for improving critical infrastructure cybersecurity. `http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf`, Feb. 2014.

[213] of State Legislatures, N. C. Security breach notification laws. `http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx`, Jan. 2015.

[214] Oliveira, R., Pei, D., Willinger, W., Zhang, B., and Zhang, L. The (in) completeness of the observed internet as-level structure. *IEEE/ACM Transactions on Networking (ToN) 18*, 1 (2010).

[215] Oracle. How do i disable java in my web browser, 2012.

[216] Parasites, U. Practical guide to dealing with Google's malware warnings. `http://www.unmaskparasites.com/malware-warning-guide/`.

[217] Path, R. The global email deliverability benchmark report, 2h 2011. `http://www.ris.org/uploadi/editor/1337589256returnpath_globaldeliverability2h11.pdf`, Mar. 2012.

[218] Picanso, K. E. Protecting information security under a uniform data breach notification law. *Fordham L. Rev. 75* (2006), 355.

[219] Pnueli, A. The temporal logic of programs. In *Foundations of Computer Science, 1977., 18th Annual Symposium on* (1977), IEEE, pp. 46–57.

[220] Povey, D. Kerberos, ldap and active directory: Single sign-on for unix and beyond. *AUUGN* (2005), 243.

[221] Provos, N., Mavrommatis, P., Rajab, M., and Monrose, F. All your iFrames point to us. In *Proc. 17th USENIX Security Symp.* (Aug. 2008).

[222] Provos, N., McNamee, D., Mavrommatis, P., Wang, K., and Modadugu, N. The ghost in the browser: Analysis of web-based malware. In *Proc. 1st USENIX Workshop on Hot Topics in Understanding Botnets (HotBots'07)* (Cambridge, MA, Apr. 2007).

[223] Qu, Z., Rastogi, V., Zhang, X., Chen, Y., Zhu, T., and Chen, Z. Autocog: Measuring the description-to-permission fidelity in android applications. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014), ACM, pp. 1354–1365.

[224] Rajab, M., Ballard, L., Jagpal, N., Mavrommatis, P., Nojiri, D., Provos, N., and Schmidt, L. Trends in circumventing web-malware detection. Tech. rep., Google, July 2011. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/papers/rajab-2011a.pdf.

[225] Ramachandran, A., and Feamster, N. Understanding the network-level behavior of spammers. In *ACM SIGCOMM* (2006), vol. 36, ACM.

[226] Ramachandran, A., Feamster, N., and Vempala, S. Filtering spam with behavioral blacklisting. In *Proc. of ACM CCS* (2007), ACM.

[227] Ramos, D. A., and Engler, D. Under-constrained symbolic execution: correctness checking for real code. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 49–64.

[228] Rao, J. M., and Reiley, D. H. The economics of spam. *The Journal of Economic Perspectives 26*, 3 (2012), 87–110.

[229] Rashid, F. Y. Grum botnet: Down one month, no impact on spam. http://www.securityweek.com/grum-botnet-down-one-month-no-impact-spam, Aug. 2012.

[230] Reaves, B., Shernan, E., Bates, A., Carter, H., and Traynor, P. Boxed out: blocking cellular interconnect bypass fraud at the network edge. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 833–848.

[231] Research, I. X.-F. Security trends in the healthcare industry. http://public.dhe.ibm.com/common/ssi/ecm/se/en/sel03048usen/SEL03048USEN.PDF?, Oct. 2015.

[232] Rinott, Y., and Tam, M. Monotone regrouping, regression, and simpson's paradox. *The American Statistician 57*, 2 (2003).

[233] Roesch, M., et al. Snort: Lightweight intrusion detection for networks. In *LISA* (1999), vol. 99, pp. 229–238.

[234] Romanosky, S., and Acquisti, A. Privacy costs and personal data protection: Economic and legal perspectives. *Berkeley Technology Law Journal 24*, 3 (2009), 1061–1101.

[235] Romanosky, S., Acquisti, A., and Sharp, R. Data breaches and identity theft: When is mandatory disclosure optimal? In *Proceedings of WEIS 2010* (2010), TPRC.

[236] Roughan, M. Simplifying the synthesis of internet traffic matrices. *ACM SIGCOMM 35*, 5 (2005).

[237] Roveta, F., Caviglia, G., Di Mario, L., Zanero, S., Maggi, F., and Ciuccarelli, P. Burn: Baring unknown rogue networks. In *Proc. of VizSec* (2011).

[238] Sanger, D. Fine line seen in u.s. spying on companies. *The New York Times. May 20* (2014).

[239] Sanger, D., and Fackler, M. Nsa breached north korean networks before sony attack, officials say. *The New York Times. January 18* (2015).

[240] Sanger, D. E., and Perlroth, N. Us said to find north korea ordered cyberattack on sony. *The New York Times 17* (2014).

[241] Schmidt, A. The estonian cyberattacks. *The fierce domain–conflicts in cyberspace 1986* (2012), 1986–2012.

[242] Schneider, F. B. Implementing fault-tolerant services using the state machine approach: A tutorial. *ACM Computing Surveys (CSUR) 22*, 4 (1990), 299–319.

[243] Schneier, B. *Applied cryptography: protocols, algorithms, and source code in C*. john wiley & sons, 2007.

[244] Schneier, B. Want to evade nsa spying? Don't connect to the Internet. *Wired* (Oct. 2013).

[245] Schulte, E., Fry, Z. P., Fast, E., Weimer, W., and Forrest, S. Software mutational robustness. *Genetic Programming and Evolvable Machines 15*, 3 (2014), 281–312.

[246] Schwarz, G., et al. Estimating the dimension of a model. *The annals of statistics 6*, 2 (1978), 461–464.

[247] Shacham, H., Page, M., Pfaff, B., Goh, E.-J., Modadugu, N., and Boneh, D. On the effectiveness of address-space randomization. In *Proceedings of the 11th ACM conference on Computer and communications security* (2004), ACM, pp. 298–307.

[248] Sharif, M. I., Lanzi, A., Giffin, J. T., and Lee, W. Impeding malware analysis using conditional code obfuscation. In *NDSS* (2008).

[249] Shirey, R. Network working group: Internet security glossary, version 2. RFC 4949, Aug. 2007.

[250] Siemborksi, R., and Melnikov, E. Smtp service extension for authentication. RFC 4954, July 2007.

[251] Sinha, S., Bailey, M., and Jahanian, F. Shades of grey: On the effectiveness of reputation-based blacklists. In *Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on* (2008), IEEE, pp. 57–64.

[252] Somayaji, A., and Forrest, S. Automated response using system-call delay. In *Usenix Security Symposium* (2000), pp. 185–197.

[253] Somayaji, A., and Forrest, S. Automated response using system-call delays. In *In Proceedings of the 9th USENIX Security Symposium* (2000), pp. 185–197.

[254] Soska, K., and Christin, N. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 33–48.

[255] Sotirakopoulos, A., Hawkey, K., and Beznosov, K. I did it because i trusted you: Challenges with the study environment biasing participant behaviours. In *SOUPS Usable Security Experiment Reports (USER) Workshop* (2010).

[256] SpamHaus. Spamhaus composite blocking list. `http://cbl.abuseat.org/totalflow.html`, May 2015.

[257] Stiennon, R. Read the mandiant apt1 report. now. *Forbes* (Feb. 2013).

[258] Stone-Gross, B., Kruegel, C., Almeroth, K., Moser, A., and Kirda, E. Fire: Finding rogue networks. In *Computer Security Applications Conference, 2009. ACSAC'09. Annual* (2009), IEEE, pp. 231–240.

[259] Stone-Gross, B., Kruegel, C., Almeroth, K., Moser, A., and Kirda, E. Fire: Finding rogue networks. In *ACSAC* (2009), IEEE.

[260] Stoughton, A. 3.17 proving the security of the mini-app private information retrieval protocol in easycrypt. *The Synergy Between Programming Languages and Cryptography* (2014), 45.

[261] Stringhini, G., Mourlanne, P., Jacob, G., Egele, M., Kruegel, C., and Vigna, G. Evilcohort: detecting communities of malicious accounts on online services. In *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 563–578.

[262] Student. The probable error of a mean. *Biometrika* (1908).

[263] Sunshine, J., Egelman, S., Almuhimedi, H., Atri, N., and Cranor, L. F. Crying wolf: an empirical study of ssl warning effectiveness. In *Proceedings of the 18th conference on USENIX security symposium* (Berkeley, CA, USA, 2009), SSYM'09, USENIX Association, pp. 399–416.

[264] Sutton, J. Gibrat's legacy. *Journal of economic literature* (1997), 40–59.

[265] Symantec. 2012 internet security threat report. `http://www.symantec.com/security_response/publications/archives.jsp`, Apr. 2013.

[266] Symantec. 2014 internet security threat report. `http://www.symantec.com/security_response/publications/archives.jsp`, Apr. 2015.

[267] Tang, Q., Linden, L. L., Quarterman, J. S., and Whinston, A. Reputation as public policy for internet security: A field study. In *Proc. ICIS* (2012).

[268] Taylor, B. Sender reputation in a large webmail service. In *CEAS* (2006).

[269] Thomas, R. C., Antkiewicz, M., Florer, P., Widup, S., and Woodyard, M. How bad is it?–a branching activity model to estimate the impact of information security breaches. *A Branching Activity Model to Estimate the Impact of Information Security Breaches (March 11, 2013)* (2013).

[270] Track, T. Majority of malware analysts aware of data breaches not disclosed by their employers. `http://www.threattracksecurity.com/press-release/majority-of-malware-analysts-aware-of-data-breaches-not-disclosed-by-their-employers.aspx`, Nov. 2013.

[271] Tran, T., Pelizzi, R., and Sekar, R. Jate: Transparent and efficient javascript confinement. In *Proceedings of the 31st Annual Computer Security Applications Conference* (2015), ACM, pp. 151–160.

[272] Traynor, I. Russia accused of unleashing cyberwar to disable estonia. *The Guardian* (May 2007).

[273] Trustwave. Trustwave 2013 global security report. `https://www2.trustwave.com/2013gsr.html`, 2013.

[274] Ugarte-Pedrero, X., Balzarotti, D., Santos, I., and Bringas, P. G. Sok: Deep packer inspection: A longitudinal study of the complexity of run-time packers. In *Security and Privacy (SP), 2015 IEEE Symposium on* (2015).

[275] Union, I. T. Ict facts and figures. Tech. rep., International Telecommunications Union, May 2014.

[276] van Eeten, M., Bauer, J. M., Asghari, H., Tabatabaie, S., and Rand, D. The role of internet service providers in botnet mitigation: An empirical analysis based on spam data. Tech. rep., OECD Publishing, 2010.

[277] Van Valen, L. A new evolutionary law. *Evolutionary theory 1* (1973), 1–30.

[278] Vasek, M., and Moore, T. Do malware reports expedite cleanup? an experimental study. In *Proceedings of the 5th USENIX conference on Cyber Security Experimentation and Test* (2012), USENIX Association, pp. 6–6.

[279] Venkataraman, S., Brumley, D., Sen, S., and Spatscheck, O. Automatically inferring the evolution of malicious activity on the internet. In *NDSS* (2013).

[280] Verizon. 2014 data breach investigations report. `http://www.verizonenterprise.com/DBIR/2014/`, 2014.

[281] Violino, B. Spam levels creep back after rustock botnet takedown. `http://www.securityweek.com/grum-botnet-down-one-month-no-impact-spam`, Apr. 2011.

[282] Visa. Visa international operating regulations summary of changes. `http://usa.visa.com/download/merchants/visa-international-operating-regulations-summary.pdf`, Oct. 2011.

[283] Voit, J. *The statistical mechanics of financial markets*. Springer Verlag, 2005.

[284] Wagner, A. K., Soumerai, S. B., Zhang, F., and Ross-Degnan, D. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of clinical pharmacy and therapeutics 27*, 4 (2002), 299–309.

[285] Walsh, S. Canadian pharmacy spam group reinvents self as "world pharmacy". *All Spammed Up* (Dec. 2010).

[286] Wang, H., Zhang, Y., Li, J., Liu, H., Yang, W., Li, B., and Gu, D. Vulnerability assessment of oauth implementations in android applications. In *Proceedings of the 31st Annual Computer Security Applications Conference* (2015), ACM, pp. 61–70.

[287] Wang, Y.-M., Beck, D., Jiang, X., Roussev, R., Verbowski, C., Chen, S., and King, S. T. Automated web patrol with strider honeymonkeys: Finding web sites that exploit browser vulnerabilities. In *NDSS* (2006).

[288] Watson, R. N., Woodruff, J., Neumann, P. G., Moore, S. W., Anderson, J., Chisnall, D., Dave, N., Davis, B., Gudka, K., Laurie, B., et al. Cheri: A hybrid capability-system architecture for scalable software compartmentalization. In *Security and Privacy (SP), 2015 IEEE Symposium on* (2015), IEEE, pp. 20–37.

[289] WhiteHat Security. Whitehat website security statistic report: Winter 2011, 11th edition. `https://www.whitehatsec.com/assets/WPstats_winter11_11th.pdf?doc=WPstats_fall10_10th`, 2011.

[290] Widup, S. The leaking vault: Five years of data breaches. *Digital Forensics Association* (2010).

[291] Wilcox, C., Papadopoulos, C., and Heidemann, J. Correlating spam activity with ip address characteristics. In *INFOCOM WKSHPS* (2010), IEEE.

[292] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics bulletin* (1945), 80–83.

[293] Williamson, M. M. Throttling viruses: Restricting propagation to defeat malicious mobile code. In *Computer Security Applications Conference, 2002. Proceedings. 18th Annual* (2002), IEEE, pp. 61–68.

[294] Williamson, M. M. Throttling viruses: Restricting propagation to defeat malicous mobile code. In *Proc. ACSAC '02* (Las Vegas, Nevada, Dec. 2002).

[295] Wolfram Research, Inc. Mathematica 10.1, 2015.

[296] Wong, C., Bielski, S., Studer, A., and Wang, C. Empirical analysis of rate limiting mechanisms. In *Recent Advances in Intrusion Detection* (2005), Springer, pp. 22–42.

[297] Wong, C., Bielski, S., Studer, A., and Wang, C. On the effectiveness of rate limiting mechanisms. In *Proc. RAID '05* (2005).

[298] Wong, C., Chan, W., and Kam, P. A student t-mixture autoregressive model with applications to heavy-tailed financial data. *Biometrika 96*, 3 (2009).

[299] Wong, C., Wang, C., Song, D., Bielski, S., and Ganger, G. R. Dynamic quarantine of internet worms. In *Dependable Systems and Networks, 2004 International Conference on* (2004), IEEE, pp. 73–82.

[300] Wong, E., and Tatlow, K. D. China seen in push to gain technology insights. *The New York Times* (June 2013).

[301] Wurzinger, P., Bilge, L., Holz, T., Goebel, J., Kruegel, C., and Kirda, E. Automatically generating models for botnet detection. In *Computer Security ESORICS 2009*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, pp. 232–249.

[302] Yang, X., Wetherall, D., and Anderson, T. A dos-limiting network architecture. In *ACM SIGCOMM Computer Communication Review* (2005), vol. 35, ACM, pp. 241–252.

[303] Yen, T.-F., and Reiter, M. K. Revisiting botnet models and their implications for takedown strategies. In *Principles of Security and Trust*. Springer, 2012.

[304] Yevstifeyev, M., Melnikov, A., and Newman, C. Pop3 over tls, 2011.

[305] Yousuf, H. 'godfather of spam' going to prison. *CNN Money* (Nov. 2009).

[306] Zaliapin, I., Kagan, Y., and Schoenberg, F. Approximating the distribution of pareto sums. *Pure and Applied Geophysics 162*, 6 (2005), 1187–1228.

[307] Zengerle, P., and Cassella, M. Millions more americans hit by government personnel data hack. *Reuters* (July 2015).

[308] Zhang, J., Durumeric, Z., Bailey, M., Liu, M., and Karir, M. On the mismanagement and maliciousness of networks. In *(to appear) Proceedings of NDSS 2014* (2013).

[309] Zhou, M., and Carin, L. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Learning* (2013).

[310] Zhou, Y., and Evans, D. Understanding and monitoring embedded web scripts. In *Security and Privacy (SP), 2015 IEEE Symposium on* (2015), IEEE, pp. 850–865.

[311] Zhu, T., Phipps, D., Pridgen, A., Crandall, J. R., and Wallach, D. S. The velocity of censorship: High fidelity detection of microblog post deletion. arxiv preprint, 2013.

[312] Zhuang, L., Dunagan, J., Simon, D. R., Wang, H. J., Osipkov, I., and Tygar, J. D. Characterizing botnets from email spam records. *LEET 8* (2008).

[313] Zou, C., Gong, W., and Towsley, D. Code red worm propagation modeling and analysis. In *Proceedings of the 9th ACM conference on Computer and communications security* (2002), ACM, pp. 138–147.

[314] lvarez, G., and Petrovic, S. A new taxonomy of web attacks suitable for efficient encoding. *Computers & Security 22*, 5 (2003), 435 – 449.