TU Delft

Delft University of Technology

Evaluating Hosting Provider Security Through Abuse Data and the Creation of Metrics

Noroozian, Arman

DOI
10.4233/uuid:8d2b0432-7ebe-42c0-b231-34f1a08bd779

Publication date
2020

Document Version
Final published version

Citation (APA)
Noroozian, A. (2020). *Evaluating Hosting Provider Security Through Abuse Data and the Creation of Metrics*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:8d2b0432-7ebe-42c0-b231-34f1a08bd779

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

EVALUATING HOSTING PROVIDER SECURITY THROUGH ABUSE DATA AND THE CREATION OF METRICS

Arman Noroozian

EVALUATING HOSTING PROVIDER SECURITY

Through Abuse Data and the Creation of Metrics

DISSERTATION

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Monday 10th of Feb. 2020 at 15:00 o'clock

by

ARMAN NOROOZIAN
MSc. in Computer Science, Delft University of Technology, The Netherlands
Born in Tehran, Iran

This dissertation has been approved by the promotor(s):

Prof.dr. M.J.G van Eeten

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof.dr. M.J.G van Eeten	Delft University of Technology, Promoter

Independent members:

Prof.dr.ir. R.L. (Inald) Lagendijk	Delft University of Technology
Prof.dr.ir. Aiko Pras	University of Twente
Prof.dr. Marianne Junger	University of Twente
Prof.dr.ir. Wouter Joosen	KU Leuven
Prof.dr. Nicolas Christin	Carnegie Mellon University

Reserve member(s):

Prof.dr.ir. Pieter van Gelder	Delft University of Technology

Other member(s):

Dr. Maciej Korczynski	University of Grenoble Alpes

This research has been funded by NWO (grant nr. 12.003/628.001.003), the National Cyber Security Center (NCSC) and SIDN, the .NL Registry.

Delft Academic Press Printed in the Netherlands by Delft Academic Press
Cover design: Shahab Zehtabchi

Distributed by Delft University of Technology, Faculty of Technology, Policy and Management, Jaffalaan 5, 2628BX Delft, the Netherlands.

ISBN 97890-6562-4451

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License, except where expressly stated otherwise.
http://creativecommons.org/licenses/by-nc-sa/3.0/

Keywords: hosting providers, cybercrime, abuse, governance, economics, metrics, incentives, hosting.

Dedicated to Saman and Maya

ACKNOWLEDGMENTS

I am certain that every path to a PhD is unique. Yet for many that undertake this journey it is commonly riddled with obstacles and set backs that one needs to overcome with numerous unforeseeable ups and downs along the journey. From having to hear that one is not PhD material, to facing the pressures of academic publishing, difficulties in attending scientific gatherings due to one's place of birth, all the way to the joy of finding out new things and contributing to science and society. This journey is as much about personal perseverance as it is about curiosity, collaboration and even at times sheer luck. But what I have mostly come to realize at this point, is that it has been such a privilege to spend several years of my life doing things that I am passionate about, to learn along the way, and to develop not only in skills but also in character.

Writing this dissertation is undoubtedly the result of several years of collaborative work. I find my self thinking that I could not have accomplished this task without the immeasurable support of many wonderful individuals whom I would like to sincerely thank.

I would especially like to thank my coauthors who have helped in conducting my studies, provided me with guidance, ideas, data and most importantly their time and friendship. These include the wonderful professors Rainer Böhme, Tyler Moore, Katsunari Yoshioka and Damon McCoy in addition to several other PhD students and Post-Docs, Geoffrey Simpson, Daisuke Makita, and Sumayah Alrwais with whom I have collaborated from afar. And closer to home, this list includes my wonderful colleagues and friends Maciej, Carlos, Michael and Samaneh in addition to Jan and Eelco from the National High Tech Crime Unit (NHTCU).

I am also greatly indebted to my dear friends and colleagues from our research team, whom in the several years that I have worked by their side, have been an inspiration. They have of course also contributed to what I am proud of accomplishing today in many ways. Whether it be through ideas, useful code and personal experiences that they have shared, guidance and their friendly banter which have all created the wonderful team that I work in. These include Rene, Hadi, Qasim, Orcun, Elsa, Rolf, Ugur, Kate, Wolter and an extensive list of new and former colleagues that have come and gone.

To my other colleagues in particular my wonderful peers from my peer group, MAS colleagues and the wider TPM faculty and those that I have interacted with through out the years I extend my warm gratitude for the wonderful environment that you have created and thank you for your support through out the years.

My dear friends Shahab, Mahtab, Nika, Pourya, Chris, Maaike, Shahab, Andreas, Dora, Shahin, Sami, Ardalan, friends from 'the island', the climbing friends, family in the Netherlands and at home, life long friends Homayoun, Ashkan, Behnam, and the many other wonderful individuals whom I have had the privilege of knowing throughout the years, I thank you for your wonderful friendship, support and shelter that you have provided to me and my family all of which have contributed directly and indirectly to my journey.

Of course none of this would have been possible without the support of my close family, my dear partner Saman, Baba, Maman, Madar, Omid, Azadeh and my extended family and family in-law whom I dearly love and thank for their unwavering support.

And finally my dear promoter Michel, whom I will save a unique spot for, and would like to thank and acknowledge for his mentoring, patience, and friendship throughout the years. It is hard to put down in words how Michel has enriched my journey. His insights, feedback, help in clarifying my manuscripts all the way to his contagious enthusiasm for scientific research and the spirit which he has injected into our team have all contributed to shaping me into to an independent researcher for which I am greatly thankful. I thank you for this wonderful journey.

CONTENTS

Part I

INTRODUCTION

1 WHY EVALUATE HOSTING PROVIDER SECURITY PRACTICES

1.1 CYBERCRIME AND THE ABUSE OF HOSTING SERVICES

Internet content is typically *hosted* on servers operated by specific intermediary businesses known as *hosting providers*. They provision servers, Internet connectivity, and storage capacity to their customers to place content online. At the moment of writing for example, I may rent a dedicated server with 4 CPU cores, 16GBs of RAM, 2TBs of storage space, along with a combined 10TBs of inbound or outbound traffic for the price of 38,00 Euros per month from the Dutch hosting company `LeaseWeb`. I may use this server to setup a personal website for my self, privately backup files or share photos with family members, or setup an online business to sell products through a web-shop hosted on the server for instance. Multi-national companies like `OVH`, `1&1`, and `GoDaddy` are all examples of hosting providers that operate in this space and provide such a service.

Our use of the Internet is largely facilitated and shaped by such types of *Internet intermediaries* ⊛ as we increasingly create, consume, and interact with digital content over the Internet through their services.

⊛ ***Internet Intermediary*** *is a term often used to refer to companies like ISPs, hosting providers, online domain registrars, online payment processors, search engines and social media platforms, to name a few, that enable and facilitate the use of the Internet [1].*

And while most hosted Internet content is benign, miscreants may also put up harmful content by abusing the infrastructure and services of hosting providers. This is of course also the case with many other types of Internet intermediary services •.

For instance, so-called *phishing* web pages put up by cybercriminals are maliciously designed to resemble the legitimate websites of our online banking or e-mail service to name a few examples. When browsed, phishing pages trick visitors into revealing their credentials or other forms of sensitive data to unintended recipients who will abuse the sensitive information if divulged.

• *Miscreants abuse various Internet intermediary services and are quite imaginative in how and what online resources they exploit. For example they misuse hosting services [2, 3], domain names [4, 5, 6, 7, 8], Domain Name System (DNS) resolution services [9, 10, 11] and mail servers [12, 13] to name a few others.*

Miscreants also host malicious code online for instance to redirect unsuspecting users to other types of malicious web pages. These may in turn employ so-called *exploit-kits* behind the scenes to infect the machines of unsuspecting visitors with other pieces of harmful code through exploiting vulnerabilities in their browser software for instance. If they succeed, miscreants can then offload among others *banking trojans*, *backdoor shells*, and *ransomware* onto user machines, which are in turn employed to steal, gain access to, or hold valuable user data hostage. Such harmful code may be broadly referred to by the encompassing term *malware*. And while some malware has the capability

to spread itself even further by automatically probing more devices for exploitable flaws, other types of hosted harmful code are instead designed to control groups of already malware-infected devices (*bots*). Bots may be directed to preform certain tasks through commands issued via a Command-and-Control (C&C) center hosted on a server, for instance to launch so-called Denial of Service (DoS) attacks against other servers to overload and knock them offline.

In short, phishing websites, malware executables, infrastructure for commanding and controlling machines that have been compromised with malware, fake online pharmaceuticals shops, underground hacker forums and markets, or even child sexual abuse material are all but a few examples of what cybercriminals host online, often with the ultimate aim of making money off of their victims [14] as a large fraction of Internet-based crime has fundamentally transformed to be driven by profit motives (c. f. Franklin et al. [15]).

Large volumes of harmful content are detected on the Internet on a daily basis [16]. Google Safe Browsing (GSB) [17] - an initiative to track and mitigate phishing and malware spreading webpages - for example reports of 1.7 million active phishing pages on Dec 15, 2019 with projections suggesting this number to be on the rise [18]. For the same time point, GSB also reports 28,000 dangerous malware spreading websites which it deems harmful. Substantial amounts of harmful content such as these typically remain unaddressed and accessible online for extended periods of time [19, 20, 21].

Hosting providers and the services they afford are a critical enabler of legitimate online activities. Yet, miscreants also abuse (or in technical jargon 'attack') hosting services, either by exploiting shortcomings in security, compromising the resources that have been provided to others for legitimate use, or by directly acquiring hosting services to criminal ends themselves [22, 23, 24]. This raises a complex question of how to deal with hosting service (in-)security. We do not clearly understand which hosting providers are abused, how often, and what role they (should) play in addressing the negative side-effects caused by the abuse of their services.

1.2 THE VARIOUS TYPES OF HOSTING

Hosting providers typically offer a diversified portfolio of services to their customers. These range from the provisioning of more expensive *dedicated* servers to relatively less expensive Virtual Private Servers (VPSs) to the even cheaper options of *shared* hosting. Dedicated hosting means that customers rent servers for exclusive use and are thus also assigned dedicated IP addresses for their servers. In shared hosting several customers share usage of the same server while also having to share the same server IP address. A VPS is a hybrid between the

latter two where infrastructure is virtually separated so it appears that customers have dedicated access while in reality they partly share server infrastructure. That is their virtual servers may operate from the same physical machine but each receive a dedicated IP address of their own to communicate with their VPS.

Hosting services may additionally include server management support (*managed* hosting) while other times they do not (*unmanaged* hosting). Especially in the case of unmanaged hosting, providers have less oversight over rented servers but then also assume less responsibility when things go wrong, for example when data backups fail or the server is lacking critical software security patches. Cheaper hosting solutions such as shared hosting typically include server management support and are managed with the help of the provider for reasons having to do with access privileges and maintaining control over certain parts of the shared infrastructure. In practice providers typically offer combinations of the aforementioned hosting solutions.

Hosting providers also come in various shapes and sizes. Larger providers typically own physical infrastructure which they locate and operate from within their own data centers. Smaller providers instead rely on ISPs or other larger providers to accommodate physical infrastructure in their data centers (so called '*collocation*'), or rely on them to provide connectivity to global networks (so called '*peering*'). Some particular hosting providers do not even own any physical infrastructure and instead 'resell' services of other providers as go-betweens through so-called 'reseller' programs (also known as *reseller* hosting). In short, depending on their business model and needs, providers may directly posses or rent small or large numbers of resources (e.g. IP addresses, servers, network infrastructure and middle-boxes) that they'll have to manage and maintain.

It is also common for hosting providers to offer other core Internet services in conjunction with their hosting solutions. Most hosting packages include domain name resolution services - a core Internet service that allows others to communicate with servers through *domain names*, for example `myserver.mydomain.com`, rather than an assigned IP addresses like `54.154.156.125`. Some also sell domain names to their customers and act as domain name registrants as well.

Extraordinarily some hosting providers are criminal undertakings. These, which are known as Bullet-Proof Hosting (BPH) providers, knowingly allow the abuse of their services. They cater to cybercriminals by for example advertising in underground markets and even offer protection against law enforcement actions to take down harmful content, thereby provisioning a stable online environment for cybercirminals to conduct illicit online activities.

In summary, variations in size and types of services offered by hosting providers, the myriad business models which they have, in addition

to the multiple jurisdictions in which they operate, give rise to a complex and heterogeneous global hosting market. The complexities of this market mean that providers are not easily and clearly distinguishable from other intermediary businesses at scale.

1.3 COMBATING ABUSE

Formal Governance of the Hosting Market

A wealth of literature on cybercrime and cybercrime business models demonstrate that almost all involve a component of abusing hosting services [24, 25]. Whether it is cybercrime involving spam emails [13, 26, 27, 28], banking fraud [3], selling of fake or illegal goods [29, 30], selling of drugs, hired guns, or other components of cybercrime in underground markets [31, 32], operating botnets [33], credential phishing [23], spreading of malware [34, 35], or even operating malicious BPH providers [36, 37]. Thus hosting providers have in theory, a pivotal role in preventing various forms of cybercrime.

So what are hosting providers legally required to do when it comes to abuse? In practice, their security practices are governed by jurisdiction-specific regulation which may be strict or more lenient depending on the region.

Within the European Union for example, hosting provider practices are governed by the *eCommerce Directive* which does not hold hosting providers liable for the misuse of their services by customers [38], as long as they are not negligent and react to legal requests to take down harmful content [39]. Similarly, within the United States, providers are not liable for harmful content as governed by the *Communications Decency Act* [40] under similar conditions.

Certain types of abuse however, for example hosting child sexual abuse material, are treated differently and providers may be held liable both within the EU and the United States for not taking action against it [41, 42] if they are aware and informed of its existence on their servers. Under the European Convention on Cybercrime for example, the creation, distribution and accessing of such material constitute criminal offenses. Other forms of content, for example adult pornography or extremist manifestos, may only be considered illegal in some jurisdictions while not in others.

Regulation largely influences and shapes the security practices of hosting providers as it sets a baseline for what providers are required to do both in terms of security practices and handling of abuse.

Status Quo Versus Best Practices

Due to their pivotal role, providers could combat abuse proactively as well as reactively [43]. For instance they could prevent compromise by patching exploitable software and support less experienced customers whose resources may be more easily compromised due to their lack of experience. They may also for example monitor their network infrastructure for signs of abuse, and suspend servers that are involved in abusive activities until assurances are gained that problems have been remediated. They could also completely take down abused resources, or clean them for future use if that is still an option. Reactions to abuse should also be quick to prevent further harm to others.

In practice, however, provider responses to abuse vary substantially [44]. In each service tier, the same contractual obligations and industry norms that determine what services are provisioned to a customer, also determine what responsibilities hosting providers and their customers have in matters of security and abuse. And there are essential differences here.

On a dedicated hosting server for instance (and to a lesser extent on a VPS), customers exert almost full control over the operating system, other software, and the content placed on servers. That is, they enjoy administrative privileges over the whole server. Unless customers request support, security responsibilities are typically shifted on to the customer even though this deviates from some of the advised security best practices. This is especially the case when talking about unmanaged hosting.

On shared hosting on the other hand, customers operate under restricted privileges on a machine they share with others. Here, customers have limited control over content and specific software which they use, and no control over operating system and other administrative server software. Thus server maintenance responsibilities, as well as those of dealing with incidents (at best) fall on both the provider and customers. For example, security conscious providers may patch and update operating system software during maintenance cycles - something that customers do not have control over in a shared hosting environment- and may additionally provide customers with patched and up-to-date versions of customer-specific software to install [45]. But then typically it is upon the customer to make that choice. Customers might not install patches as they could break the functionality of software which they use. On the other hand providers with lax security practices, may not even provide patches to their customers for various reasons including that it is costly to do so [46].

Regardless of which of these scenarios plays out, hosting providers' pivotal role in preventing abuse is undeniable. They often control, hand out, and operate the underlying resources that either point to

content, host content, or run code. If and when these resources are abused, they are in key positions to monitor for, or respond to various manifestations of abusing these resources [43]. Yet expecting providers to actually fulfill such a role would be going beyond base line regulatory requirements.

Luckily, within the hosting market, certain 'soft' forms of governing have emerged from the industry itself as attempts to move providers beyond baseline requirements set by regulation. These are attempts to get providers to implement more effective countermeasures against abuse. Among them, the Messaging Malware Mobile Anti-Abuse Working Group (M3AAWG) - a respected global industry initiative to combat harmful content - sets forth an proposes a number of security best practices for hosting providers to follow [47]. Table 1.1, as published within the most recent version of these guidelines, illustrates several types of hosting services, highlights the parties that are normally in control of various resources, in addition to propose which parties should be responsible for dealing with the abuse of resources. M3AAWG's guidelines clearly go beyond regulatory requirements. Providers are advised for example to take responsibility by blocking or removing harmful content proactively, in addition to reactively if and when informed of abuse. Such recommendations are much more in line with what hosting providers could theoretically do against abuse.

Table 1.1: Various types of hosting with respect to parties that control resources and proposed reponsibilities for dealing with abuse issues as best practice

Hosting Type	Hardware	Operating System	Software	Abuse Issues
Dedicated	Provider	Customer	Customer	Customer
Managed	Provider	Provider	Provider	Provider or Customer
Reseller	Provider or Customer	Customer or its Client	Customer or its Client	Customer or its Client
Shared	Provider	Provider	Provider	Provider and Customer
Unmanaged	Provider and Customer	Customer	Customer	Customer
Virtual Private Server	Provider	Provider	Customer	Customer

Albeit that such best practices are steps towards the right direction within the hosting market, following self-regulatory norms are voluntary of course. Current governance structures have, for now, proven to be ineffective in addressing the problem of abuse as evidenced by the large volumes of it that remain unaddressed globally. Current regulation has translated to many providers shifting security responsibilities to customers or other third parties in practice [48]. Thus, Internet intermediary responsibilities towards preventing the abuse of their infrastructure, and by definition also that of hosting providers, are exceedingly a topic of discussion among academics and regulators [42, 43, 49].

As matters stand, hosting providers are understood to mostly take voluntarily action against harmful content hosted via their infrastructure with some being more vigilant than others. Reputation effects and peer pressure may act as a form of incentive for voluntary adherence to additional security practices such as the ones proposed by M3AAWG. However, if hosting providers are to move beyond baseline regulatory requirements, creating the right incentives to adhere to stronger security practices is clearly a critical problem that needs solving. Therefore, an exceedingly important question, one with which this work is concerned with, is how hosting providers could be incentivized to do more against abuse.

1.4 GOVERNANCE CHALLENGES

Collective Action and the Weakest Link Problem

Cybercrime has become a global phenomenon and dealing with it requires collective action by multiple entities to address its negative side-effects [50].

Yet, not all hosting providers implement suitable countermeasures or take action when their resources are abused [51]. The lax security practices of some providers results in a *whack-a-mole* game in which criminals are able to migrate their abusive practices and content to those lax providers even when others are vigilant and enforce suitable security countermeasures [39]. In other words, this creates a *weakest-link* problem. It appears that there is no shortage of hosting services with weak security to choose from within the global hosting market. So called Bullet-Proof Hosting (BPH) providers that are in the business of enabling cybercrime are a particularly difficult problem to tackle in this respect [52].

Given the status quo, combating abuse currently also depends, for a large part, on the security efforts of third parties to notify hosting providers of abuse and to get them to act against abuse [53, 54, 55]. The alternative is to protect Internet users by other means when providers do not, for example by taking away and blocking domain names that point to harmful content [5, 6]. Many of these efforts, notwithstanding their limitations, rely on sharing of so-called abuse data collected and disseminated by various independent parties [25]. Organizations like `Google`, `Spamhaus`, and `Shadowserver`, to name some examples, routinely monitor websites and other Internet resources for harmful content and notify various parties to take action against them. By partnering with such organizations, popular web-browsers (*e.g.* `Chrome`, `Firefox` and `Safari`) display warnings to users before they put themselves at risk by interacting with harmful content on the web. Additionally, by leveraging such abuse data, email services and

client software reject spam messages or emails that are suspected of containing harmful attachments and links to phishing websites. Some domain registries and registrars also suspend domain names that are misused towards spreading harmful content by leveraging the same kind of data [4]. Numerous third party system have also been proposed to proactively prevent the abuse of Internet resoruces or protect users against compromise (c. f. [4, 56, 57, 58, 59]), sometimes even predicting abuse before they are compromised.

All too often however, even third-party security efforts fail to get those that are in key positions to address abuse [19, 60, 61]. When all else fails, we rely on court orders and law enforcement bodies to combat cybercrime and take down harmful content or abused network resources.

In April 2018 for example, law enforcement authorities from the Netherlands, UK and US dismantled a popular website (`WebStresser.org`). It allowed any paying individual to kick ("*boot*") other Internet users or websites offline at the click of a button [62]. This so-called "*booter*" website was able to launch Distributed Denial of Service (DDoS) attacks against any victim of choice by abusing vulnerable unpatched network devices. Only a month later, Dutch and Thai police, arrested two individuals who misused rented servers and network infrastructure to operate a bullet-proof hosting business (`MaxiDed`) [63]. Its operators obtained hosting resources by entering into reseller relationships with several parenting (upstream) hosting providers. Yet, both of these examples are cases of abuse incidents that could have been prevented, by for instance adhering to M3AAWG's best security practice guidelines. In the former case, the hosting provider could have taken the booter website offline, while in the latter case, the parenting hosting providers that entered into reseller relationships with the BPH provider operators could have terminated `MaxiDed`'s reseller contracts.

As matters stand, we lack scalable countermeasures to the global problem of cybercrime (c. f. [64]). Many hosting providers do not effectively combat abuse, third-party efforts fail too often, and our last resort options are costly, and even more difficult to scale due to factors like jurisdictional complexities [39]. Addressing this shortcoming is not only a matter of technical solutions but also a matter of economic incentives [65, 66, 67], which I will discuss next.

Miss-aligned Incentives, Externalities, and a Market for Lemons

Hosting providers, like many other software-based businesses, are economically driven by such factors as network effects, and dominance within the context of economic markets. Assuming a market perspective, the security of the products and services that are sold, or their privacy implications for that matter, are not found high on the agenda

of most digital business [68]. Moreover, the existing regulation governing the hosting market which I discussed earlier, does not incentivize market players to take effective mitigatory actions against abuse. That is because they are not generally liable if and when abuse of their services takes place [43] in addition to their adherence to best security practices such as M3AAWG's being voluntary.

As a result, for hosting providers, incentives to counter cybercrime are often misaligned with the aforementioned driving economic factors. More attention is being payed to the latter than to security efforts which are typically treated as less necessary additional costs. A lack of liability for the abuse of their services has in fact been one of the driving factors behind the growth of many intermediaries' services [42].

As such, the hosting market exhibits a so-called '*market failure*' [49] with consequential negative outcomes of the kinds previously illustrated through several examples. Market failures especially occur when the negative side-effects and costs of negligence are 'externalized', or in other words borne by third parties, leading to so-called *negative externalities*® [69]. For hosting providers, the cost of cybercrime which is enabled by the abuse of their services, is borne by the individual victims, other businesses, or society as a whole [70] and typically not by themselves [71]. The law enforcement operations to take down so-called booter websites or `Maxided`'s BPH business discussed earlier are clear examples of how governments bare part of the cost. In other cases, the costs are directly borne by the victims, or may alternatively be borne by insurance companies, or for example banks who reimburse victims when their money gets stolen as the result of online banking fraud for example.

®*A* ***Negative Externality*** *is a cost born by a third-party as the result of an economic transaction*

To incentivize providers to act more responsibly and effectively against abuse, we need to be able to identify which providers perform poorly and which perform well in terms of security. Without this knowledge, the market cannot reward secure practices, nor can governance mechanisms 'internalize' the cost of abuse onto providers. In other words to make the providers themselves bare the cost of cybercrime.

But with respect to the hosting market and its failures, we lack even the most basic information such as which providers operate within the market. To the best of my knowledge, there is no technical data that clearly identifies hosting providers globally. And the data sources that may be employed for this purpose are limited, some challenging to utilize [72], difficult to parse [73], or rife with inaccuracies [74, 75]. Thus, a necessary step is to develop measurement techniques to identify and construct a global list of hosting providers from existing data before we can begin to understand which providers are more secure and which less and thus problematic. This important knowledge gap creates a situation, which in economic terms is referred to as a '*market for lemons*'. The term refers to situations in which good or

bad products (e. g. lemons) are indistinguishable due to *information asymmetry* about the quality of the products. The absence of empirical data about hosting providers, which also exists in other market areas [49], leads to information asymmetry regarding the security of hosting providers. In the hosting market context, this term refers to the fact that while hosting providers themselves inherently possess greater knowledge about their own security, other stakeholders do not.

Combined, information asymmetry and miss-aligned security incentives, exacerbate market failure problems and lead to a corrosion of incentives to combat cybercrime since we are unable to distinguish good and bad hosting providers. It has become all too common to see cybercrime as someone else's problem or something to be dealt with at a later point in time across a wide range of digital products and services [76].

While the concepts that I discuss here provide us with a theoretical economic understanding of why security in the hosting market fails, it is still a matter of researching which technical solutions and/or governance strategies are better suited for aligning the security incentives of hosting providers with their economic driving incentives. In other words, incentive schemes have to be designed such that security aspects are taken into account by hosting providers and in economic terms for cybercrime costs to be 'internalized'.

1.5 TOWARDS POTENTIAL SOLUTIONS

To address abuse, security best practices and literature call for hosting providers to, proactively patch vulnerabilities to prevent compromise (c. f. [77, 78]), implement security controls (c. f. [79]) in addition to automated solutions to monitor for abuse (c. f. [51]), and to block and remove harmful content post-haste (c. f. [43, 53]). It has been suggested that providers may also need to implement stricter policies about how and with whom they do business (c. f. [47, 52]).

But before any of these solutions are likely to be adopted across the market, we first need to fix the underlying incentive problems. And decades of experience from closely related industries with similar problems, for example the telecommunication industry, has demonstrated that fixing market failures, may also require regulators to step in, and implement suitable governance strategies to restore and realign economic incentives [80, 81, 82]. With respect to the hosting market, similar non-technical solutions may also be required [38, 49, 65, 83].

Knowledge Gap: Which hosting providers and services are most abused and insecure

The facts of the matter however are, that due to existing information asymmetry we do not clearly understand which of these solutions is going to have an effect on the hosting market nor how effective they may be. At a more basic level, the inherent information asymmetry barriers are even preventing us from distinguishing between good,

negligent, or even outright bad hosting providers, let alone empirically measure which security solutions may be more effective. In other words, even basic questions like, who are the worst hosting providers, how effective are their current security efforts, and how do these efforts compare to those of their counterparts are currently difficult to answer as we lack empirical tools to measure security outcomes within this market. Only once we can empirically measure security and distinguish failures from successes, are we going to be able to understand which solutions are better at moving the hosting market forward. For example, by empirically tracking and comparing the progress resulting from the adoption of proposed technical or non-technical solutions. As such, information asymmetry is the central problem that this dissertation attempts to tackle and reduce.

***Problem Definition**: Reducing information asymmetry about the security of hosting providers*

To this end, this thesis explores ways to empirically measure and compare the security efforts of hosting providers and their postures towards the abuse of their services as a basis for answering the fundamental questions that were posed above. I propose to design and operationalize '*security metrics*' as a possible way of measuring, monitoring, and comparing the security of hosting providers. These metrics would have to translate the information available on the abuse of hosting services into numbers that meaningfully reflect hosting providers' security postures. A crude and simplified version of this can be thought of as a scoring or ranking system. So-called abuse (or incident) data collected and disseminated by third parties are examples of externally available information that can be used as inputs. Abuse data from `Google`, `Spamhaus` and `Shadowserver` which were discussed earlier are more concrete examples.

***Possible Solution**: Design security metrics that reliably translate external information on the abuse of hosting services into numbers representing the effectiveness of provider security efforts to prevent and combat abuse*

Security metrics may potentially be employed as benchmarks to compare the effectiveness of provider security efforts, thus allowing various stakeholders to understand why, if, and where most of the abuse takes place within the hosting market. Providers themselves for instance, may employ metrics to compare their efforts against competitors, or understand which security practices are most effective, as well as track progress. Policy analysts, may similarly employ metrics to measure and empirically test the effectiveness of certain security practices at a market level thus informing and enabling regulators to enact policies and uphold standards that have measurable and demonstrable effects which are grounded in empirical data. Similarly, law enforcement agencies may use metrics to focus their efforts and pursue the worst actors. Moreover, when or if security metrics become common knowledge, consumers or other businesses can make informed choices about which hosting providers to transact with. Reputation effects may also be instrumentalized to induce competition and incentivize providers to combat abuse more effectively from a market perspective for example through insurance premiums adjusted to provider security levels.

That being said, the main focus of my work is on the design of meaningful security metrics as a primary step. The question of how to employ the metrics themselves to steer and govern the hosting market towards more desirable outcomes is something that still needs to be explored in future ongoing work which I will later discuss in Chapter 7.

1.6 STATE OF THE ART

In light of the existing market failures that I have discussed, it should come as no surprise that empirical studies of hosting provider security practices typically find their abuse mitigation efforts to be inadequate. A small scale study by Canali et al. [51] for instance clearly demonstrates that hosting providers are unable to effectively detect and block illicit activities taking place on their servers and that some common sense best security practices like running network monitoring tools are sometimes neglected. Prior to this, Christin et al. had also found evidence of disproportionate misuse of certain hosting provider infrastructure, relative to their market share, in committing certain illicit activities namely so-called one-click fraud [84]. Such empirical studies clearly point to the ineffective security practices of certain providers.

Larger scale empirical studies (c. f. [82]), including that of my own and colleagues (c. f. [44, 78]) also suggest that there are large variations in how effectively abuse is dealt with across networks, whether they be that of hosting providers, ISPs, or other types of networks.

Variations in responses to abuse, i. e. the fact that some providers are much more effective at curbing abuse than others, have been linked to security efforts by proxy of network hygiene indicators. For instance better management of infrastructure and servers empirically correlate with decreased levels of abuse [73, 85]. The reverse also holds. That is, abuse tends to concentrate around mismanaged networks implying that increased security effort may lower abuse levels.

The amount of resources that providers manage/maintain or the number of their customers, which are possible measures of the size of the hosting business, are also a key influencing factor. Measures of provider size quantify the potential attack surface of a provider and strongly correlate with abuse levels. These are highly predictive of the number of security incidents that may occur, a phenomenon that is repeatedly observed in the literature and in practice [86, 87, 88]. Naturally then, comparing providers in terms of security efforts would have to take such exposure effects into account to allow apple-to-apple comparisons. As larger providers are more likely to experience incidents in absolute numbers simply due to their larger size, it would be misleading to compare their security efforts against smaller less exposed providers.

Beyond mismanagement of infrastructure and exposure effects like provider size, several risk factors for compromise have also been highlighted in the literature (c. f. [2, 78, 89, 90]). Certain types of hosting services for instance - e. g. shared hosting or the provisioning of Content Management System (CMS) platforms like `Wordpress`, `Joomla` and `Drupal` - have been shown to elevate the likelihood of abuse. Such additional risk factors also influence provider exposure and have to also be taken into account in the design of metrics as some concentrations of abuse may be explainable by such factors rather than weak security practices.

At the same time the literature also points to additional factors, e. g., biases and errors in empirical abuse data, that contribute to patterns of elevated abuse that may at times be considered spurious [91]. And while elevated or concentrated abuse patterns around certain networks are often interpreted as indicators of bad security practices or even outright malice, biases and errors in measurements have to also be taken into account before inferences about bad security practices are drawn.

A particular area of research that has received a lot of attention is the special case of Bullet-Proof Hosting (BPH) providers, which directly cater their services to cybercriminals. Several systems have been developed to detect BPH providers (c. f. [36, 37, 92, 93]) based on symptomatic indicators such as high concentration of abuse, so-called fast-fluxing of IPs, and temporal characteristics of responses to abuse complaints. Yet, the BPH problem remains a difficult problem to solve within the hosting market as its operators adapt to evade such detection techniques.

These developments in detecting malicious networks, as well as the identification of factors that drive abuse, lay much needed groundwork in understanding the factors that influence hosting provider security outcomes, albeit that studies in this area are typically focused on certain areas of the hosting market, e. g. shared hosting, or BPH.

As such there have been recent calls for undertaking more empirical measurements and development of reliable metrics that cover broader market sections [49, 94]. Several recent studies have produced metrics for closely related Internet intermediaries such as ISPs [80] and Top-Level Domain-name (TLD) operators [11] for example. Yet the development of such metrics for hosting providers are a relatively less explored endeavor.

Limited industry and security vendor initiatives exist to produce empirical security metrics for hosting providers (c. f. [95]). Providers of blocklists and abuse data also typically produce crude metrics which count the number of abuse incidents at various levels, for instance, around IP addresses and networks (c. f. the statistical abuse reports of initiatives like `Google Safe Browsing`, `Shadowserver`, `bgpranking`, and `abuse.ch` to name a few examples). Forgoing that the commercial

security industry may have incentives to exaggerate security failures [49], the produced metrics have several drawbacks in the sense that they often do not account for endogenous or exogenous factors that shape the overall security outcomes for various hosting providers which have been identified in the literature. For example, they do not account for the well-known fact that larger providers are probabilistically more prone to their servers being misused nor the findings that cheaper shared-hosting services increase the risk of abuse. As such, they lead to biased comparisons which typically paint larger providers as negligent. Moreover, the methodologies by which these metrics are produced are opaque and thereby limit their adoptability by larger audiences. Here, of course a balance needs to be struck to prevent the metrics from being gamed.

A particular challenge in developing unbiased empirical metrics for hosting providers is that security, and hence security performance, is a dynamic multi-causal phenomena driven by a multitude of factors that are difficult to measure and disentangle [96, 97]. While some factors such as provider exposure are relatively more straightforward to take into account, others factors, for example how customers or attackers behave, and what type of harmful content ends up being hosted on their servers are not. It goes without saying that for example certain content is more harmful than others and treated differently from a legal perspective, for instance botnet C&C centers versus illegal video streaming websites versus child sexual abuse material versus hate speech). As providers treat different types of abuse differently their security performance is also affected by what their priorities are in dealing which each type of abuse. Therefore reliable metrics need to also take the types of harmful content into account in order to allow meaningful comparisons to be drawn between hosting providers.

Also challenging is the consequential fact that quantifying the security of hosting providers and thereafter making meaningful security performance comparisons require data on many aspects of their business which do not readily exist. For example, and as I have mentioned before, there is no straightforward way of globally identifying hosting providers as there is no maintained list of hosting providers that one could refer to. Businesses that provide hosting services are typically identified from ill-maintained Internet operations data such as WHOIS information and Border Gateway Protocol (BGP) data which contain Autonomous System Number (ASN) information as identifiers of organizations (c. f. [98, 99]). On the other hand, security incident data, otherwise referred to as abuse data, which is our primary source of empirical information on how the security efforts of providers manifest, is also usually limited and riddled with its own biases that are not well understood. Therefore, understanding why certain hosting providers experience more security incidents in comparison to others

and controlling for factors that are not directly under their control is a key step in making unbiased security performance comparisons.

1.7 RESEARCH AIMS

Given what I have argued, the main aim of this dissertation is to understand how the security performance of hosting providers can be reliably measured through the design of security metrics and to what extent. In other words its main question can be formulated as follows:

> `How can we quantify the effectiveness of hosting provider security practices?`

To successfully answer this research question the following sub-questions need to be answered to form a coherent understanding of cybercriminal misuse of hosting provider infrastructure and the possible applications of security metrics to mitigate this problem:

- **RQ1**: What steps are required to translate empirical abuse data into meaningful security metrics for hosting providers such that they reliably quantify and signal the effectiveness of their security practices relative to other hosting providers?
- **RQ2**: How can we infer the proactive security performance of hosting providers (relative to others) from noisy abuse data?
- **RQ3**: How we can we quantify the reactive security performance of hosting providers (relative to others) from noisy abuse data?
- **RQ4**: Are security metrics effective in identifying criminal Bullet-Proof Hosting (BPH) providers and, if not, how does BPH operate and why do security metrics fail? Moreover, what alternative pressure points can we find to disrupt their operations?
- **RQ5**: How do lax security practices translate into wider societal problems and what are the wider effects of the cybercrime that it (un)wittingly facilitates?

The answer to each of the more focused research questions outlined above brings us closer to forming a better understanding of the cybercriminal misuse of hosting provider services, in addition to how and for which particular circumstances security performance metrics are a useful solution.

1.8 DISSERTATION OUTLINE

The remainder of this dissertation is structured according to the outline presented in Table 1.2. Each chapter directly corresponds to the research questions outlined earlier in corresponding order. This table also

provides information about the publications on which each chapter is based.

Table 1.2: Dissertation Outline

Chapters	Research Question	Based on Publications
Chapter 2	RQ1	Arman Noroozian, Maciej Korczynski, Samaneh Tajalizadehkhoob, and Michel van Eeten. "Developing Security Reputation Metrics for Hosting Providers." In: *USENIX CSET*. 2015
Chapter 3	RQ2	Arman Noroozian, Michael Ciere, Maciej Korczynski, Samaneh Tajalizadehkhoob, and Michel Van Eeten. "Inferring the Security Performance of Providers from Noisy and Heterogenous Abuse Datasets." In: *WEIS*. 2017
Chapter 4	RQ3	Arman Noroozian, Geoffrey Simpson, Maciej Korczynski, Tyler Moore, Rainer Bohme, and Michel van Eeten. "Using Abuse Data to Evaluate Remediation Efforts." 2018 (yet to be published)
Chapter 5	RQ4	Arman Noroozian, Jan Koenders, Eelco van Veldhuizen, Carlos Hernandez Ganan, Sumayah Alrwais, Damon McCoy, and Michel van Eeten. "Platforms in Everything: Analyzing Ground-Truth Data on the Anatomy and Economics of Bullet Proof Hosting." In: *Proc. of Usenix Security Symposium*. 2019
Chapter 6	RQ5	Arman Noroozian, Maciej Korczyński, Carlos Hernandez Gañan, Daisuke Makita, Katsunari Yoshioka, and Michel Van Eeten. "Who Gets the Boot? Analyzing Victimization by DDoS-as-a-Service." In: *Proc. of RAID*. 2016
Chapter 7	Main RQ	Discussion and conclusions based on all publications listed above.

Given the relatively unexplored state of security performance metrics for hosting providers, Chapter 2 investigates existing security performance metrics for hosting providers and takes a broad look at what information about hosting providers is required to construct meaningful security metrics. This chapter also explores what steps need to be taken to translate available and relevant information into security performance metrics for hosting providers. The chapter set the agenda and maps what subsequent steps to take to answer my main research question. Next, Chapter 3 investigates how the proactive security efforts of providers can be externally measured and how to deal with the inherent noisy nature of abuse data. Subsequently, Chapter 4 investigates how hosting providers react when incidents occur and how well they perform when notified of security incidents. The chapter constructs additional security metrics to compare reactive security performances of hosting providers. Next, Chapter 5 takes a closer look at the special case of criminal Bullet-Proof Hosting (BPH) providers, how they operate and whether these can be identified through security performance metrics. In Chapter 6, I step back and examine the negative side-effects of provider negligence by studying the victims of cybercrime in a case-study of Distributed Denial of Service (DDoS) attacks which are facilitated in part by negligent hosting providers that host booter websites. Finally, Chapter 7 brings together my results and discusses the implications of my findings along with my concluding remarks. Complementary material to this thesis, such as co-authorship contributions to each study are provided thereafter.

Part II

PEER-REVIEWED STUDIES

2 DEVELOPING SECURITY METRICS FOR HOSTING PROVIDERS

At the onset of my studies, existing metrics for comparing hosting provider security postures typically counted and compared instances of abuse among providers. Some of the metric outcomes demonstrated unusually high concentration of abuse at certain hosting providers. Research into cybercrime which often points to concentrations of abuse, implicitly implies that providers with high concentration of abuse are worse in terms of security; some are considered 'bad' or even 'bullet-proof' hosting providers. Concentration of abuse is also often taken to point at cases that are amendable to intervention. Yet, more recent research argues that not all concentrations should be interpreted as such, since some may be spurious and driven by data artifacts and measurement errors. Moreover, only in some cases did existing metrics take into account the differences among providers in terms of susceptibility to abuse. For example by normalizing incident counts against the size of the advertised IP address space of the provider. In other words, these other metrics compared provider security based on the number of abuse incidents per provider IP as a way of accounting for differences in exposure among providers. Remarkably though, little work existed at the time on more systematically comparing the security postures of different hosting providers.

Comparing provider security through metrics involves methodological as well as metric design choices which have an impact on the metric outcomes. And the previous attempts to compare provider security through metrics had not systematically considered such design choices, nor fully addressed some of the methodological challenges of metrics design. For instance that quantifying a provider's attack surface through the proxy of its advertised IP space is just one way of characterizing its exposure. Or for example, the fact that other factors than just exposure also drive abuse. How attackers behave, the types of hosting services [89], or the quality of abuse data [91], also impact our observations of abuse.

*Thus in this chapter I first present a systematic approach for metrics development and identify some of its main challenges: (**i**) identification of providers, (**ii**) abuse data coverage and quality, (**iii**) taking exposure into account (also referred to as 'normalization'), (**iv**) metric aggregation and (**v**) metric interpretation. I describe a pragmatic approach to deal with some of these challenges and subsequently improve on the process and metrics that I develop, later in Chapter 3 and 4.*

This chapter is based on the first of a series of peer-reviewed studies that I have conducted on this subject [100]. In the process of this study, I also answer an urgent question posed to us by the Dutch police at the time: 'Which are the most abused providers in our jurisdiction?'. Notwithstanding their limitations, there was and still is a clear need for security metrics for hosting providers in the fight against cybercrime.

2.1 INTRODUCTION

Hosting providers are companies that provide servers via which customers can make content or services available on the Internet *e.g.* websites, email or support for multi-player gaming. As with virtually all services on the Internet, they are abused for criminal purposes as well. A wealth of research has identified how hosting infrastructure shows up in various criminal business models. Think of phishing sites, Command-and-Control (C&C) servers for botnets, distribution of child sexual abuse material, malware distribution, and spam servers (c. f. [24, 36, 52]).

Nobody contests that hosting providers play a key role in fighting cybercrime. Much of the criminal activity runs on compromised servers of legitimate customers, some on servers rented by the criminals themselves. In either case, the hosting providers typically becomes aware of the problem only after being notified of the abuse. And their response to abuse reports varies widely, ranging from vigilant to slow to negligent or even bullet-proof [36, 51, 37]. To empirically measure which of these responses is actually occurring has proven to be very challenging. Existing metrics of hosting provider security typically count instances of abuse within an Autonomous System (AS), sometimes normalized by the size of the advertised address space [36, 92, 95] to somewhat account for provider exposure ⊛. None of these attempts adequately account for the serious methodological challenges plaguing such metrics.

⊛ *Also see statistical abuse reports of initiatives like* `Google Safe Browsing`, `Shadowserver`, `bgpranking`, *and* `abuse.ch` *to name a few other examples*

In this chapter, I present a systematic approach for developing metrics for hosting providers. It enables us to identify and discuss the main challenges: (**i**) identification of providers, (***ii***) abuse data coverage and quality, (**iii**) normalization, (**iv**) aggregation and (**v**) metric interpretation in light of the heterogeneity of hosting providers. Additionally I present a pragmatic approach to deal with these issues.

This study was originally part of an ongoing collaboration with the Dutch National High Tech Crime Police, the Authority for Consumers and Markets, the Public Prosecutor and the Dutch Hosting Provider Association. Its objective is to answer an urgent question posed by the police at the time: *'which are the worst providers in our jurisdiction'*?

The question in itself illustrates that there is a clear need for security metrics for hosting providers, notwithstanding their limitations of course. Reducing cybercrime is as much a problem of incentives as it is a

technical issue [67]. Without reliable metrics to signal provider security we cannot tell which provider is vigilant, lax, negligent or outright criminal and it will be very difficult to move the sector towards more secure practices. Here, and as I have discussed before in the previous chapter, *Information asymmetry* erodes the incentives of providers to invest in security. Reliable metrics can (**i**) signal security performance to customers, upstream and downstream providers, law enforcement and other stakeholders, (**ii**) enable benchmarking of providers, and (**iii**) help identify the effectiveness of security practices and policies.

The main contributions of this chapter are as follows: (**i**) I outline a systematic process to develop security metrics for hosting providers, as well as the methodological challenges encountered along the way, (**ii**) I improve existing techniques for mapping abuse to hosting providers and for taking into account the size of hosting providers in computing metric scores, and (**iii**) I present a pragmatic approach to produce metrics for the Dutch hosting market, that was developed in collaboration with some of the key stakeholders.

2.2 BACKGROUND

Hosting providers come in many shapes and sizes and offer portfolios of services: from relatively expensive dedicated physical machines to virtual private servers (VPS) to the cheaper options of shared hosting or even so-called free hosting. In each service, the role of the provider vis a vis the customer is different. On a dedicated machine, and to a lesser extent on a VPS, the customer controls the entire software stack, whereas on shared hosting, many customers operate under restricted privileges on a machine they share with many other users. Free hosting services limit user control to the extreme.

Depending on the type of customer, hosting providers play a different role in protecting their customers against compromise by patching servers, cleaning, and monitoring for abuse. Similarly, providers need to protect the rest of the Internet against potentially malicious customers by putting in place different checks and restrictions which depends on the service contract with that customer.

Next to the rate at which abuse incidents occur, the remediation time (which I will also refer to as 'uptime') of abuse, also reflects hosting provider security practices. On one end of the spectrum, vigilant hosting providers remove malicious content often within hours of its discovery, in the middle there are some providers that respond more slowly and more selectively, and on the other extreme are the so called Bullet-Proof Hosting (BPH) hosting providers that seem to ignore all abuse notifications.

There has been a lot of speculation over the security incentives of providers. A shared hosting provider, for example, could act against

abuse more directly because its customers have only limited control over the machines that they use. On the other hand, shared hosting is a highly competitive market with low margins, so investing in security is not likely to be a high priority. The only way forward is to replace speculation with reliable empirical evidence of abuse rates across providers based on data.

2.3 OVERVIEW OF APPROACH

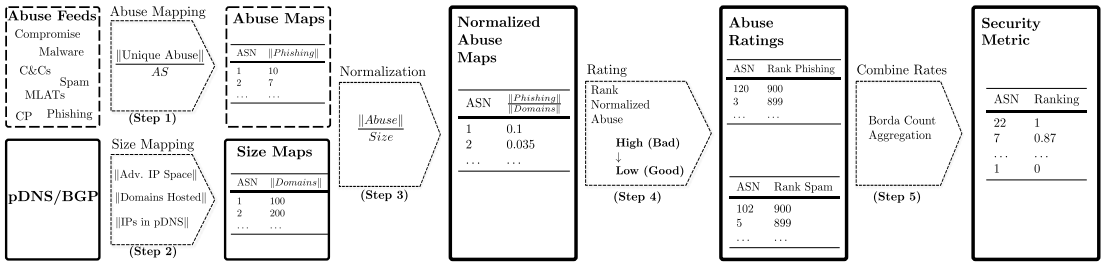


Figure 2.1: Illustrative overview of security metric development process. Starting form the left, a series of processing transformations (represented as arrows) are applied to data artifacts (represented by rectangular boxes) to produce security metrics for comparing hosting provider security. Each graph element contains illustrative examples of the type of data (illustrated as data tables) and the transformations that are applied to produce the next step.

My proposed approach for calculating hosting provider security metrics is partly guided by the goal to allow our collaborators to engage in meaningful discussions based on reliable empirical techniques. To this end, I produce two types of security indicators for hosting providers based on data available in public and private abuse feeds: (***i***) *Occurrence of abuse*: an indicator based on counting occurrences of abuse incidents, and (***ii***) *Persistence of abuse*: an indicator based on how long the abuse was present and how long it took for the incident to be remediated. Distinguishing between the occurrence of abuse and the remediation response of a hosting provider to abuse incidents as independent measures of performance is important. While the occurrence of abuse is to some extent inevitable due to technical vulnerabilities and related to organization size and attacker characteristics, persistence of abuse indicates attitude towards dealing with abuse and mainly relates to defender characteristics. In conjunction, these independent indicators provide a better understanding of the overall security performance of a hosting provider.

Figure 2.1 illustrates a high-level overview of the complete procedure to produce these indicators. Here, boxes represent data artifacts as inputs and outputs of each step while arrows transformations that are applied to the data. The process is generic and outlines the steps that

any security metric requires to arrive at final scores for the security of hosting providers. In executing these steps there are challenges that need to be overcome and choices that have to be made that will undoubtedly effect the reliability and interpretation of the metric. In what follows, I systematically walk the reader through these steps meanwhile highlighting challenges related to each and the possible effects they have on the overall metric and its interpretation. A more detailed analysis of some choices and their effect are presented later in Section 2.10.

2.4 STEP 1 - ABUSE MAPPING

Identifying hosting providers is not straight forward since they do not directly map onto entities with which underlying Internet protocols work or what abuse data capture. The first decision that needs to be made thus is to *identify what a hosting provider is*.

Identifying Hosting Providers

To produce security metrics for Dutch hosting providers, I have made the (common [36, 52, 92, 105]) practical assumption that hosting providers will have an associated Autonomous System Number (ASN). Consequently, I initially consider any AS which routed IP addresses geolocating to the Netherlands as a Dutch hosting provider ⊛.

⊛ *This was done using Maxmind's commercial geo-location data which is known to have inaccuracies [106]. Note however, that this step was only done to limit the scope of the study. The process of overall process of producing metrics described in the previous section does not rely on this step*

While the assumption may hold in general, ASes (and their associated ASNs) may refer to Internet Service Providers (ISPs), Internet exchange points, banks, governmental institutions, universities, and in general non-hosting entities as well. Without a deeper analysis of the ASes, such an assumption may lead to considerable error in mapping abuse onto hosting providers. Even when an AS does refer to a hosting provider, further complexity still exists. Certain providers may have multiple ASNs, or there may be multiple organizations which own a smaller part of the IP space routed from within an AS, e. g. contain *reseller* hosting providers who lease infrastructure from the AS owner. Some ASNs also advertise ranges and route traffic destined to and from IPs owned by their *peers*. Furthermore, certain legitimate services (*e.g.* CloudFlare) may act as proxies and hide the true providers hosting certain IP ranges. As a result abuse associated with small organizations with registered IPs within ASes may end up attributed to the AS from which the infrastructure is leased. Typically, the aforementioned simplifying assumption to identify hosting providers needs to be balanced out with requirements of the metric and whether the abuse from each of these smaller organizations needs to be taken into account. Here however, I

have opted to treat abuse for smaller organizations with registered IP ranges inside larger ASes in later chapters.

One method to better identify hosting providers and the potential organizations under each AS, is to analyze IP '*ownership*' using Maxmind's GeoIP ISP Database [107]. Utilizing such information results in a more fine grained mapping which mitigates the mapping problems discussed above. Nevertheless, this approach has complications of its own such as non-standardized WHOIS data formats where the same organization might appear with multiple names that are non-trivial to relate to each other. For instance, the Dutch hosting provider `Leaseweb` also appears under the following additional names in WHOIS: `Leaseweb Asia Pacific. ltd.`, `leaseweb1.iomadserve.com`.

Unit of Abuse

The second key decision is about the *unit of abuse* or *how to count the abuse data*. Unlike other hosting metrics which typically count distinct IP addresses as the unit for abuse (c. f. [36, 52, 92]), my proposed approach considers unique 2nd-level domain-IP pairs - $\langle 2LD, IP \rangle$ - as the unit of counting abuse. From this point on in the chapter, I use the terms '2LD' and 'domain' interchangeably, unless the context requires otherwise. Simply counting the number of abusive IP addresses largely underestimates abuse from shared hosting services since criminals may use the same IPs for various purposes. For example a compromised server may host a phishing website and also be used for spreading malware. Furthermore, the number of domains is a better proxy for the number of customers of the provider, which is valuable to include in approximating its size. Last, this definition also maximizes the value of our feeds as measured by their differential contribution [108].

Counting pairs of $\langle 2LD, IP \rangle$ mitigates the problem but is not perfect. In some cases it is appropriate to count fully qualified domain name and IP pairs $\langle FQDN, IP \rangle$ pairs (e.g. malicious domain generation algorithms), or even $\langle URL, IP \rangle$ pairs (e.g. child sexual abuse content concentrated under the same domain with varying paths in the URL).

Data feeds

A separate decision in mapping abuse is *what data feeds to use*. A wide range of abuse on the Internet is associated with hosting. Hosts are used as malware drop zones and to host phishing pages designed to steal sensitive information. Botnet Command-and-Control (C&C) servers are also hosted [52]. Other types of hosting related abuse includes child sexual abuse material, illicit Search Engine Optimzation (SEO) schemes, spam and counterfeit goods stores. Not all criminal activity can be

observed in a way that can be attributed to the infrastructure of a specific hosting provider however. Think of hidden services on `Tor`. Even if it can be observed, the criminal activity might not be captured in abuse data feeds, which are often produced by automated means. This implies that abuse feeds are always partial and of varying quality. This is a well known fact [108, 109, 110]. Needless to say, criminal activity that is not captured in the abuse data included in a metric, forms a blind spot of that metric. This suggests to include as broad a spectrum of abuse feeds as possible [108].

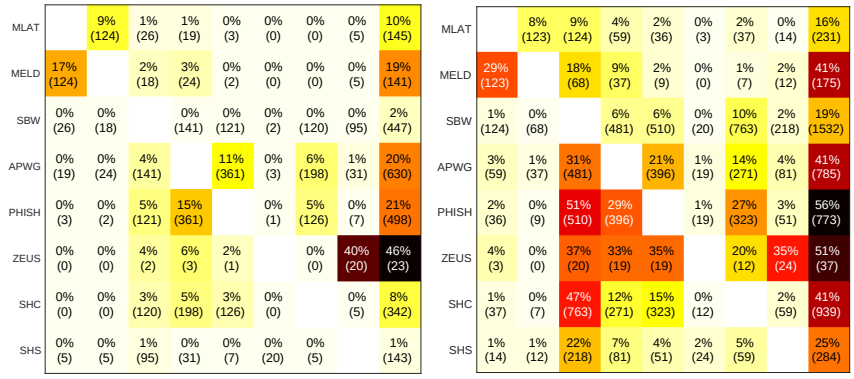
For the purpose of this study I collected a range of feeds and blocklist data from private, public, commercial, and governmental sources. Table 2.1 gives and overview of these data feeds. The data spans over the entire duration of 2014 (with the exception of the SHC and SHS which span over the 2nd half of 2014). The majority of the feeds, do not share much information on the exact collection methodology. I did not include some of the available spam feeds because our analysis of their data revealed these to be mostly related to compromised end-user machines residing within ISPs rather than hosting companies. In general, data quality relates mainly (but not only) to: (i) coverage (*What is the overlap between the different feeds?*), (ii) purity (*How much of the flagged domains truly host malicious content?* However, it is not always possible to assess the coverage or purity of a feed since many of its details are not well documented [108].

Coverage. Previous overlap analysis of blocklists capturing different types of abuse concludes that - although existent - there is little overlap in terms of the abuse associated with each ASN [109]. I have reached similar conclusions especially when $\langle 2LD, IP \rangle$ pairs are the unit of counting abuse.

Table 2.1: Statistics on collected abuse feed data employed to construct the security metrics discussed in this chapter

Abuse Type	**Feed**	**Organization**	**Samples**			
			$\langle Domain, IP \rangle$		IPs	
			Total	Excl.	Total	Excl.
Malicious Hosts	SHC	Shadowserver	3,957	3,615	2,260	1,321
Malicious Hosts	SHS	Shadowserver	7,632	7,489	1,100	816
Malware	SBW	StopBadware	15,204	14,757	7,702	6,170
Botnet C&Cs	ZEUS	Abuse.ch	50	27	72	35
Phishing	PHISH	Phishtank	2,278	1,780	1,377	-
Phishing	APWG	APWG	3,060	2,430	1,886	1,101
Take Down Request	MLAT	Dutch Police	1,347	1,202	1,433	1,202
Child Pornography	MELD	Meldpunt	725	584	417	242
Total			34,253	31,884	16,247	11,491

Clearly the feeds differ substantially in terms of the volume of reported abuse samples. For example, the professionally sourced *SBW* feed contributed over 15,000 samples, while the non-profit *ZEUS* feed three orders of magnitude less domain-IP pairs. In terms of the total number of IPs, *SBW* reports almost two times less unique IPs than distinct $\langle 2LD, IP \rangle$ pairs whereas the *ZEUS* feed reports more IPs because some ZEUS malware config, binary, and drop zones are hosted solely on IP addresses not associated with domains. Moreover, the differences between $\langle 2LD, IP \rangle$ pairs and IPs indicate that many domains used for criminal activity are mapped to a smaller number of IP addresses which could be the result of shared hosting services. Across the abuse feeds, 93% of all $\langle 2LD, IP \rangle$ pairs and 71% of all IPs for *all domains* were exclusive to a single feed (cf. *Excl.* column in Table 2.1). I refer to samples as *exclusive* when they appear only in one feed.



(a) $\langle 2LD, IP \rangle$ overlap in feeds (b) IPs overlap in feeds

Figure 2.2: Pairwise overlap of feeds with unique $\langle 2LD, IP \rangle$ and IPs as unit of abuse. $(n) =$ `number of items in intersection`.

Figure 2.2 illustrates pairwise feed intersections as a matrix, with unique $\langle 2LD, IP \rangle$ (a) and unique IPs (b) as the unit of abuse respectively. Here darker color shades represent larger overlaps among the data in the compared feeds. For instance, in Figure 2.2 (a), the overlap between *MLAT* and *MELD* data, indicates 124 $\langle 2LD, IP \rangle$ pairs in common. This overlap constitutes 9% of the MLAT feed. In comparison, 124 $\langle 2LD, IP \rangle$ pairs represents 17% of the MELD feed. The rightmost column of the matrices indicates the absolute number and the percentage of samples that the feed has in common with all other feeds combined. The amount of overlap in the rightmost columns confirms the simultaneous use of IPs for different malicious purposes. This also further supports our choice of abuse unit. In other words there is more unique information in each feed when considering $\langle 2LD, IP \rangle$ as the unit of abuse. This may

be observed by the overall smaller overlap of each feed with all other feeds combined, in comparison to the overlaps when considering the alternative choice which is depicted on the right side of the figure.

Finally, the relatively small overlap among our chosen data feeds in terms of $\langle 2LD, IP \rangle$ indicates the suitability of these feeds, nevertheless, other feed characteristics still need to be analyzed to further establish suitability.

Purity. All abuse feeds contain false positives. The main question is which samples should be considered false positives and excluded and which should not? I define false positives as websites maintained by legitimate users that do not serve any malicious content and are incorrectly flagged.

Additionally, some domains are legitimate but merely point to servers that host malicious content. For example, I consider URL shortening services such as `goo.gl` or `bit.ly` as false positives. However, other legitimate websites such as free web hosting providers (e.g. Hostinger), or cloud storage services (e.g. Imagezilla.net or Dropbox) are actually misused by criminals and included in the analysis.

Moreover, a certain portion of abuse feeds include benign domains. I analyze benign domains that appear in the Alexa top 25 thousand domain list to evaluate the prevalence of false positives in the collected data. Although I did not undertake real-time verification of flagged Alexa domains appearing in the data, I perform *a posteriori* analysis to further establish the suitability and quality of the data and the feeds themselves.

For brevity however, I only briefly discuss the analysis and do not include the details. Overall I find a limited number of Alexa domains in the abuse feeds. Nevertheless, there are major differences among the types of Alexa domains per feed. For example, through manual analysis of a random sample from the *SHS* feed I found that approximately 30% of this feed's Alexa domains were file sharing services most probably used to host malicious content and thus relevant to include. On the other hand, I also find some examples of popular websites like `msn.com`, or `microsoft.com` that are presumably used by compromised machines to test network connectivity. In the case of the *PHISH* feed I found a significant number of ranked domains that are either false positives (e.g. banks and other legitimate services), or are not appropriate for the type of analysis that we provide (e.g. URL shorteners). The majority of the ranked domains for both *MLAT* and *MELD* represent file and adult content sharing services. As systematic false positives, or unrelated web services in the feeds do not constitute a large fraction I therefore have opted to include ranked domains in my analyses.

2.5 STEP 2 - SIZE MAPPING

Reliable security metrics need to account for a commonly observed trend that larger providers also experience a larger amount of abuse.

A common yardstick for measuring the 'size' of hosting providers is the number of IP addresses routed by its corresponding AS in the BGP protocol [36, 92]. Nevertheless not all IPs routed by an AS are used for hosting content nor are they directly in use by the AS. IPs may be leased and used for other purposes. Inaccuracies in size estimation may negatively impact the reliability of a metric in that they can lead to misleading results. Nevertheless, due to simplicity of calculating, advertised IP space remains and attractive choice for size estimation.

I propose (and use) two additional size estimators: (***i***) the *number of hosted 2LDs* and (***ii***) the *number of IP addresses used to host content* per hosting provider. To calculate these estimators I use historical passive DNS (pDNS) data provided by Farsight Security [111]. This data records DNS resolution queries collected over the entire duration of 2014 which I use to count the number of unique 2LDs and their matching IP addresses. These counts are subsequently mapped to ASes routing the IP addresses and used as an estimation of the size of the provider. Here the quality of the resulting estimates is highly dependent on the coverage of the pDNS data. As long as the pDNS data has a reasonable coverage of all registered 2LDs, it can be used to produce reasonable size estimates. Hence I have crosschecked the number of unique 2LDs observed in the pDNS data with the number of 2LDs of Generic Top-Level Domains (gTLDs) and Country Code Top-Level Domain (ccTLD) present in zone files of new gTLDs that were obtained under agreement from ICANN ⊛ in addition to ccTLD sizes reported by APWG • at the end of 2014. Extrapolating from these results I have concluded pDNS to be a reasonably reliable source to estimate hosting provider size (see [112]).

⊛ http://newgtlds.icann.org

• http://docs.apwg.org

There are potentially other conceivable estimators for hosting provider size such as the number of customers. Nevertheless, the scarcity of data to base such estimates on is a largely limiting factor in this respect.

2.6 STEP 3 - NORMALIZATION OF ABUSE

Given the output of the abuse mapping and size mapping steps, the next step in the metric production process is to normalize abuse volume by a size estimate. This leads to $S \times N$ normalized abuse mappings where S is the number of size maps produced earlier and N the total number of abuse maps corresponding to the analyzed abuse feeds. A key question here relates to *interpretations that can already be made from normalized abuse data*.

All size estimates have their advantages and disadvantages which have to be viewed as trade-offs. The most commonly used size estimator

- routed IPs - is the easiest to calculate, but it suffers from systematically favoring large providers, since not all routed IPs are used for hosting. Using the portion of the routed IP space that is used for hosting as the size estimator mitigates the problem, however, this is much more difficult to calculate. This estimate is also not free of systematic bias, because it favors hosting providers that have a disproportionately large amount of shared hosting. We can use the number of hosted 2LDs as the estimator, which would treat shared hosting fairly but would still underestimate the size of subdomain resellers and free-hosting services. The trend here is clear; normalized abuse has its blind spots, and needs to be taken into account especially for interpreting results at this stage.

It is important to note that some size estimates are more volatile than others due to the dynamic nature of the underlying processes. For example, the number of FQDNs hosted by a provider may change at a much faster rate than the number of 2LDs if an estimator based on FQDNs is used.

Normalized abuse, is already an indicator of security performance by itself. Note however, that normalized abuse is abuse-type specific. For example, one can analyze normalized abuse based on the occurrence of malware on hosting providers and draw conclusions; however, this only provides a partial picture of the performance of hosting providers. Some providers might be much less strict about allowing malware spread from their servers than for example the hosting of child sexual abuse material [52]. In my case, I use all size estimators outlined in the previous section without committing to a specific one or considering one superior to others. The expectation is that the combination of these can overcome the deficiencies of each. This matter is further explored in Section 2.10.

Finally, note that when talking about metrics based on the uptime of abuse, size corrections are not appropriate. In such cases it is common to use mean or median uptimes instead of normalized abuse.

2.7 STEP 4 - RATING OF ABUSE

Given the normalized abuse data, the next step in the process calculates rankings over all maps to produce rankings. Rankings are one way of unifying the scales on which normalized abuse is measured and allows cross comparisons over categories of abuse. For example, comparing the security performance of a hosting provider in terms of how well it manages to mitigate malware with its performance in terms of how well it mitigates phishing is not meaningful when based on normalized abuse. However, the comparison is meaningful over rankings.

Given the normalized abuse maps, my method for ranking hosting providers is as follows: I rank normalized abuse from high to low. This results in $3 \times N$ *rankings*. The individual rankings may range between

zero and R, the total number hosting providers being ranked. The worst rank, R, is assigned to the AS (i. e. provider) with the highest normalized abuse, $R-1$ to the second worst and so forth. ASes with equal normalized abuse are assigned equal ranks. If a normalized abuse map only contains data on for example 20 providers the ranking will range between $[R-20, R]$ with all providers for which no abuse was detected receiving the low rank of $R-20$.

An important consideration in producing rankings is *information loss*. To illustrate this consider two hosting providers HP_1 and HP_2 that have a normalized abuse of 0.1 and 0.3 and have been assigned the ranks of 10 (worst performer) and 5 (5th worst) respectively. Ranking is not *distance preserving* since the difference between the hosting provider ranks $(10-5)$ does not entail the same information as that of the normalized abuse $(0.3-0.1)$. That is, one unit of change in ranking could mean any number of changes in the unit of normalized abuse. As a result these distances cannot be interpreted in the same way. In ranking hosting providers, some information about the magnitude of the differences is unavoidably lost.

2.8 STEP 5 - AGGREGATION OF RATES

I now aggregate the previously constructed rankings into one overall ranking that assigns scores in the range $[0, 1]$, where score 1 indicates the worst performer. The aggregation procedure considers every ranking as a voting preference over R candidates in an imaginary election. The election winner is effectively decided using a *Borda Count* vote aggregation method that basically counts how many times a certain candidate appeared in the 1st place, 2nd place, 3rd place (and so forth) in every ranking and decides the outcome based on all rankings.

An alternative approach could perform factor analysis and take into account the most contributing feeds when interpreting metric scores (cf. also Section 2.10). We find, however, voting systems to be a useful analogy when thinking about aggregation. A useful aggregation method must have certain desirable properties, such as being intuitive. For example, if a particular hosting provider is the worst ranked performer in all categories of abuse, the security metric should reflect that by assigning the worst metric score to that provider and not to others. Certain methods of aggregation will not guarantee such properties and are therefore undesirable. I refer the reader to literature on different voting aggregation methods (c. f. Cranor's discussion on vote aggregation methods ⊛ in [113]) for a better understanding of the properties of such methods and their limitations.

⊛ http://lorrie.cranor.org/pubs/diss/node4.html

2.9 STEP 6 - METRIC INTERPRETATION

Security metrics need to be interpreted to guide policy and reduce information asymmetry around the security performance of hosting providers. However, correct interpretation of a metric without detailed knowledge of the various blind spots and biases of the process is difficult. Additionally the heterogeneity in the hosting provider landscape directly influences what conclusions can be drawn from the metrics.

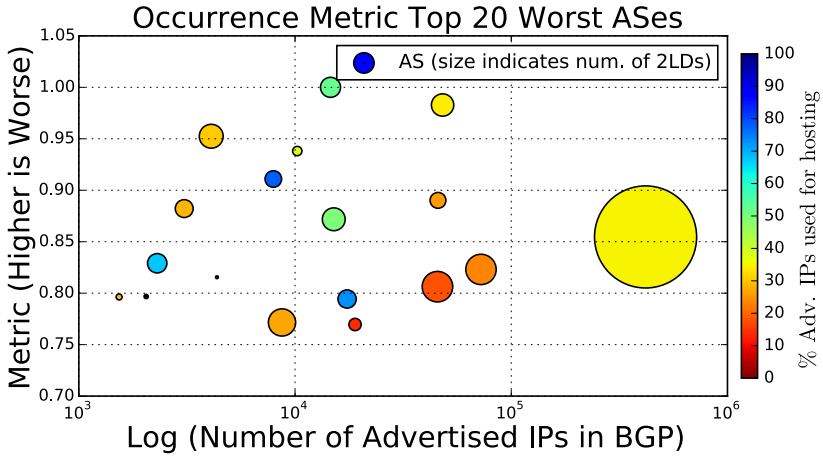


Figure 2.3: The 20 worst Dutch providers for compared by abuse rate. Security metric scores based on the occurrence of abuse are reported on the y-axis. Provider exposure to abuse is reported along the x-axis, as well is encoded in color and bubble size. Larger bubbles and shades leaning towards blue colors indicate larger exposure while smaller circles and color shades approaching red indicate reduced exposure to abuse.

To illustrate the challenges of interpretation I briefly present some of our results here. Figure 2.3 plots the metric rankings of the 20 worst identified Dutch hosting providers based on the occurrence of abuse. The plot demonstrates a large variance between the security performance of providers that have comparable size. The results clearly indicate significant differences in how hosting providers deal with abuse. Here, the safest comparisons are among providers that have the most similar properties. As an example consider the two hosting providers colored in bright green, first the provider with the highest (worst) metric score, and second, the provider located approximately at $(x = 10^4, y = 0.85)$. These data points represent providers very similar in all exposure aspects and therefore it can be safely concluded that the provider with the lower score is performing significantly better than the worst performer due to its security policies and practices. To consider the worst provider as negligent or criminally engaged simply

because it has the worst score is however a wrong conclusion to draw here.

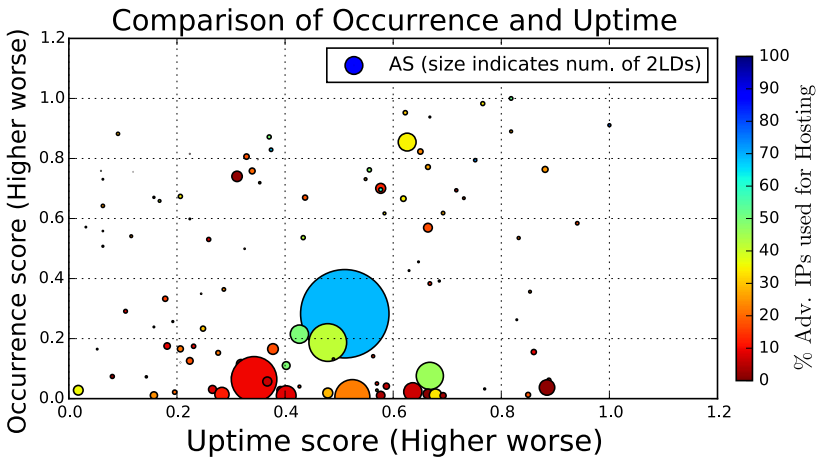


Figure 2.4: Comparison of occurrence and uptime metric for Dutch hosting providers. Provider security metric scores based on the occurrence of abuse are reported along the y-axis. Metric scores based on the remediation response of providers (uptime) are reported along the x-axis. Measures of provider exposure to abuse are color coded as well as reported through bubble size.

Alternatively, Figure 2.4 compares the occurrence metric and the uptime metric of all the identified Dutch hosting providers. A cautionary note here is that the uptime metrics are based on only 2 data feeds from which incident remediation times could be calculated. The weak relationship (Spearman's $\rho = 0.38$, Pearson $r = 0.36$) between occurrence and uptime is expected as each captures a different aspect of hosting provider characteristics that relate to abuse (see Section 2.3). Clearly some providers experience large amounts of abuse while managing to quickly block the abuse (upper left region of the plot). Others, perform consistently bad in the sense that they experience large volumes of abuse relative to their exposure and are also slow to block it (upper right region of the plot). Nevertheless, we believe that the amount of occurring abuse and the response of a hosting provider to abuse are important aspects that need to be measured separately to provide a thorough picture of security performance. Only now can we draw the conclusion that the worst performing hosting provider in terms of abuse occurrence is probably negligent because it is also among the worst performers in terms of remediation (see point with $(x = 0.8, y = 1)$ coordinates).

Finally when interpreting the results, one should also take the hosting provider business model into account. Hosting providers with a

large portion of shared hosting customers have a larger role to play in cleaning up content than ones with dedicated hosting clients. It might very well be the case that the worst performer in terms of both occurrence of abuse and response to abuse provides solely unmanaged hosting to its customers and therefore not in the same position as its peers that provide mainly shared hosting for instance. In this case the observed performance could simply be indicating the security of the hosting customers rather than that of the hosting provider itself.

2.10 SENSITIVITY ANALYSIS

To better understand the impact of key metric design decisions, I undertake a brief sensitivity analysis of alternative specifications for *(i)* unit of abuse, *(ii)* abuse normalization, and *(iii)* metric aggregation strategies. I explore the robustness of the results by producing rankings based on alternate methodological options for each of these decisions. I compare them to the ranking of the pragmatic approach that I presented (referred to as benchmark ranking) by calculating the *Pearson correlation coefficient* among the top $n = 100$ worst performing Dutch hosting providers.

Unit of abuse. In the benchmark ranking I used unique $\langle 2LD, IP \rangle$ pairs as the unit of abuse. I also calculated an alternate ranking based on *unique IP counts*, the standard approach in the literature. The Pearson's r among these alternative rankings is $0,952$. Perhaps more tangible: 16 ASes are in the top-20 of worst performers of both rankings. In other words, the metric is not very sensitive to either specification.

Abuse normalization. I also calculated three alternate scores using the three size estimates for hosting providers: *(i)* the advertised IP space *(ii)* the advertised IP space that is used for hosting, and *(iii)* the number of hosted 2LDs. The benchmark ranking used all three of them. The Pearson's r for the alternative specifications to the benchmark ranking are $0,896$, $0,909$, and $0,6438$, respectively. These results reveal a strong correlation between the IP space-based size estimators and the benchmark ranking, and a less strong correlation with the estimates based on 2LDs. Out of the top 20 worst performers in the benchmark ranking, only 7 ASes were present in all alternate top 20 rankings. When comparing pairwise: the reference ranking shares 13 ASes in with those based on advertised IP addresses and hosted 2LDs, and 15 ASes with the ranking based on IP addresses utilized for webhosting. In other words, using the number of hosted domains vs. estimates based on IP address space as measure of exposure give significantly different results. By including all three size estimations, the benchmark metric specification mitigates that impact, while retaining the advantages of including domain name counts in abuse counting and size estimation.

Metric aggregation. I also compare the benchmark ranking with a ranking in which I assign weights to each data source according to its comprehensiveness, i.e., the relative volume of each feed in terms of distinct number of exclusive $\langle 2LD, IP \rangle$ pairs in the dataset (c. f. Table 2.1). In summary, out of top 20 abused ASes in the benchmark, 14 ASes showed up in the weighted rankings and the Pearson's r of the benchmark ranking and weighted ranking is equal to 0,791.

2.11 RELATED WORK

Numerous studies have pointed to concentrations of abuse in certain networks, typically in the context of a specific criminal business model, e.g., spam [80], phishing [60] or malware [114]. Effects and policy implications of intervention at classes of intermediaries have also been studied in [7, 80]. The quality of abuse data has also been extensively covered [108, 109, 110, 114]. [51] examines the role of hosting providers in detecting abuse and reacting to user complaints for shared hosting providers. It paints a general picture that underlines the need for hosting security metrics. None of these studies try to develop security metrics from abuse data, however.

Closest to this study are the studies [36, 92, 115]. [105] produces a weighted metric score for ASes. [36] includes only up-time data and focuses mainly on identifying the worst actors. Here I expand on this work by systematically addressing challenges not discussed there. These studies typically count IP addresses as the unit of abuse and use advertised IP address space as a normalization factor.

Industry attention to hosting has been along the lines of the Host Exploit Index (HE index) [95]. While valuable, its methodologies are not fully transparent and the parts that are, suffer from similar limitations as the academic work discussed above.

2.12 CONCLUSIONS

In this chapter I have systematically worked through some of the many challenges of developing security metrics for hosting providers. All conceivable metrics will suffer from various limitations, that much is clear. This is not to say that they are not useful. When we presented this approach to various stakeholders in the Netherlands, including hosting providers, the main response was that the metrics were a valid starting point for evaluating hosting security, incentivizing self-regulation and, ultimately, identifying actors for enforcement activities.

As long as the heterogeneity of hosting providers and methodological limitations that I discuss are taken into account, valid inferences may be drawn from the metrics developed in this chapter to compare the

security of hosting providers. That is both in terms of the prevalence of abuse as well as provider reactions to abuse incidents. Inferences may for example be drawn to measure market responses to interventions, or to asses the effectiveness of security policies with the aim of steering the hosting market towards more desirable security outcomes. These metrics allow stakeholders to gain insight and empirical grounding in support of such efforts. However, the limitations that I have discussed also clearly indicate that the way forward is to improve this methodology in order to allow valid inferences to be drawn more easily.

This can be achieved by first, including and better combining abuse the data captured from multiple feeds, which I will discuss next in Chapter 3, as well as include more uptime data in constructing metrics, which I discuss in Chapter 4. In what follows I will also undertake more in-depth sensitivity analysis of how the various methodological factors impact metric outcomes, as well as improve the methodology for identifying hosting providers by employing WHOIS data on IP address ownership, rather than AS-level routing data. While the more in-depth analysis will allow more robust inferences to be drawn from the metrics, the latter technical improvement will allow smaller hosting providers, e. g. resellers to be benchmarked as well. Last but not least in Chapter 5 I will also investigate incentives under certain conditions where security metrics can be gamed by providers.

In conclusion, It is safe to say that a lot of the current claims about hosting providers are based on anecdotal evidence or methods that are not adequately understood. This initial chapter contributes to remediating this shortfall.

3

EVALUATING HOSTING PROVIDER PROACTIVE SECURITY EFFORTS

*In Chapter 2, I systematically worked through some of the challenges of developing security metrics for hosting providers and developed two types of metrics: (**i**): metrics based on how frequently abuse incidents occur, and (**ii**) metrics based on how timely incidents are remediated. It is clear that all proposed metrics have their limitations. But that is not to say that they do not project useful information.*

Some noteworthy limitations of this previous work are the fact that I had to make simplifying assumptions about how to identify hosting providers, as well as how to account for differences among causal factors that drive the abuse of their services, for instance their exposure to abuse and attacker behavior. We also saw volatility in metric outcomes depending on how the metrics were constructed, i. e. with respect to certain design choices that were made. This may be problematic especially since the metrics that I have discussed are essentially point-estimates that do not somehow reflect these choices, the underlying uncertainties, or the level of confidence that we may instill in their outcomes.

These limitations are invariably linked to the quality and availability of empirical data which may be used to construct metrics. For example, direct observations of attacker behavior is virtually non-existent and the approach to identifying providers by their ASNs in Border Gateway Protocol (BGP) data carries limitations. With respect to some of these limitation however, the employed abuse data itself offers a work around to some of these issues. Empirical abuse data simultaneously reflect the security efforts of defenders as well as reflect how attackers behave. As such, abuse data can play an important role in understanding the drivers and causes of insecurity which can lead to solutions for strengthening and aligning the security incentives of hosting providers.

Using abuse data to measure security performance suffers from a number of problems, however. Abuse data is notoriously noisy, highly heterogeneous, often incomplete, biased, and driven by a multitude of causal factors that are hard to disentangle. In this chapter, I present a comprehensive approach to measure defender security performance from a combination of heterogeneous abuse data sets, taking all of these issues into account. I discuss a causal model of incidents and on its basis propose a data modeling approach which employs Item Response Theory (IRT) towards estimating provider security as a latent trait reflected within the abuse data.

This alternative approach to developing metrics improves upon the previous chapter in several ways. First, it allows us to quantify the underlying uncertainties of the abuse data and to reflect those in the security performance estimates and metric outcomes. Despite the uncertainties, I demonstrate the effectiveness of the approach by using the security performance estimates to predict the incident frequencies observed in independent datasets, after controlling for various exposure effects such as the size and business type of the providers. Second, I employ a more reliable technique for identifying hosting providers as well as quantifying their exposure within this new study, which is based on WHOIS IP ownership information rather than Autonomous System (AS) ownership embedded within BGP routing data which I previously employed. Finally, I also verify that the simplifying assumptions regarding attacker behavior dynamics that I implicitly made in my previous work is indeed a reasonable one to make at the global hosting market level.

This chapter is based on my second study on the subject of metrics development [101], which focuses on the question of how to develop security metrics for hosting providers that reflect their ***proactive*** *security efforts.*

3.1 INTRODUCTION

Empirical observations of the computing resources that are being abused by criminals, also known as abuse data, are an important foundation for the research on cybercrime. Abuse datasets typically focus on a specific type of criminal resource – e.g., phishing sites, compromised domains, Command-and-Control (C&C) servers, or infected end user machines – depending on the automated tools via which the data is collected, such as spam traps, honeypot networks, botnet sinkholes, webcrawlers, sandboxes, and the like.

Studies based on abuse data have often looked at concentrations of incidents in certain networks [36], Internet Service Providers (ISPs) [80, 116], countries [117, 118], organizations [73], payment providers [13], registrars [5], registries [112], and other agents. The idea is that such concentrations are amenable to intervention. They are interpreted to reveal attacker economics, such as scale advantages, or defender economics, such as lack of security investment by some agents because the cost of incidents is externalized to others [91, 97].

Abuse data offers one of the very few empirical measurements of the security performance of defenders. As such, it can play an important role in strengthening and aligning the security incentives in a variety of markets. It has been used to reduce information asymmetry and leverage reputation effects [95, 119], to identify bad providers [100, 120], and to study the effectiveness of countermeasures [110, 121].

Using abuse data to measure defender security performance suffers from a number of problems, however. First of all, abuse data is notoriously noisy. It contains all kinds of issues around false positives and

negatives, incorrect attribution to the responsible agent, inconsistent measurement over time, dynamic attacker behavior, and more, which others have discussed previously (c. f. [108]) and I have also alluded to in Chapter 2.

Second, abuse datasets are highly heterogeneous. They are very different in size. Some sets observe one or two orders of magnitude more events than others. In the previous chapter, this was the case for example when comparing some of the abuse feeds that I employed. Pitsillidis et al. have similarly observed significant differences in examining the coverage of spam related abuse feeds for instance [108]. Metcalf and Spring have observed the same phenomenon across a wider range of data feeds also referred to as blocklists [109]. Moreover, they observed that abuse feeds also typically have very little overlap. Even datasets of the same type of abuse, say phishing, rarely independently observe the same incident. The correlation of different datasets can be quite low, when counting the number of incidents per defender (e.g., hosting provider). Some providers might be more susceptible to certain types of abuse, but less to others.

A third problem is the lack of completeness [108]. Not all abuse events are observed. Those that are observed might contain biases. Related to this is the fact that not all providers are observed in abuse data. All studies that start with the abuse data itself to evaluate providers will, therefore, suffer from selection bias, as providers where no incidents were observed are excluded, even though they might be performing better than those that are included.

Fourth, and final, is the problem of multicausality. Abuse data is driven by a variety of factors and it is difficult to isolate the defender's performance from them. It is clear, for example, that defenders with more infrastructure and customers will incur more incidents [97, 100]. Unless the other factors are explicitly modeled, any analysis is at risk of incorrectly assuming that differences in abuse rates reflect differences in defender efforts.

The first two problems imply that using a single abuse data source to measure defender efforts is highly unreliable, as the outcomes will differ greatly per data source. Different sources will have to be combined to derive a more trustworthy signal [108]. This also means that measurement errors have to be carefully considered as they may result in observing spurious concentrations of abuse [91]. The third problem, lack of completeness, means that sources of bias in the data have to be investigated and mitigated. One key requirement is that any analysis will have to identify the relevant market players independently from the abuse data, in order to avoid selection bias. The fourth issue, multicausality, has to be tackled by embedding any analysis into an explicit causal framework that captures, at least analytically, all the relevant forces that influence the abuse rates.

Recent work in this area has addressed one, sometimes two or three, of these problems, but no study has addressed all four. I will discuss this in more detail in the section on related work. In this chapter I present the first comprehensive approach to measure defender security performance from a combination of heterogeneous abuse datasets. I apply the approach to the hosting sector, which is associated with a large portion of all observed abuse events. I will first present a causal model to explain abuse rates in provider networks. I then map the providers in the hosting market. Second, I study potential biases in the distribution of abuse data over providers. Next, I collect relevant exposure variables for the providers. We can then specify a model, based on Item Response Theory (IRT), to estimate the security performance of providers as a latent variable from a collection of abuse datasets, while controlling for exposure effects, such as the size of the network of the provider. Last, I test the reliability of the performance metric.

The contributions of this chapter may be stated as follows:

- I formalize a causal model in order to systematically disentangle the different factors at work in abuse data. This model provides a basis for modeling security economics questions based on incident data.
- I show that a combination of 7 abuse datasets cover observations in just 34% of all providers in the hosting market. While most providers have no observed incidents, there is no evidence of bias. Via a simulation, I demonstrate that all providers, small and large, have equal probability of showing up in abuse data, once we control for their exposure.
- Next, I present a novel statistical approach – based on Item Response Theory (IRT) – to estimate the security performance of providers as a latent factor from a range of heterogeneous abuse data sources, while controlling for exposure effects.
- Finally, I demonstrate the reliability of the new performance metric. Notwithstanding the noisy nature of abuse data, using the latent variable we are able to explain between 75-99% of the variance in any independent abuse dataset, after controlling for exposure effects. This result agrees with previous work in that combining abuse data from various sources lead to a better characterization and understanding of security performance [108].

The overall goal of this study is to enable better measurement of security performance from combined sources of abuse data, while controlling for differences in firms and their exposure to attacks. The result is a security benchmark that helps to reduce information asymmetry in these markets, thus improving the security incentives of providers.

Reliable performance metrics are also critical to study impact of interventions and recommended security practices. The success of a wide range of industry and government-backed initiatives to combat cybercrime critically depend on benchmarks to provide empirical evidence through which the success and progress of the initiatives can be tracked.

In what follows I will first discuss the causal abuse framework which forms the background of this work in Section 3.2. I then provide an overview of our data in Section 3.3. To explore the bias in our abuse data, i independently map the hosting provider market using several other data sources in Section 3.4 and find no evidence of observation bias in abuse data using simple simulations of attacks across the hosting market in Section 3.5. I then move on to construct an IRT model and motivate our approach in Section 3.6, then provide the specification of the model in Section 3.7 and estimate the security performance of the hosting providers in Section 3.8. The robustness and predictive power of our security performance estimates are explored in Section 3.9. Finally I provide an overview of the related work, and studies on which this chapter builds in Section 3.10 and finally discuss the implications and conclude in Section 3.11.

3.2 CAUSAL MODEL

A lot of empirical research is based on the distribution of abuse across networks or other units of analysis. Any interpretation of those distributions makes assumptions, often implicitly, of the underlying factors at work. This is even more clear for causal inferences. Several studies looked at the relationship between characteristics of organizations, networks or providers and their abuse rates, e.g., indicators of network mismanagement [85], provider properties and business models [97], or the effect of interventions [117, 121].

Previous work shows that the variance in abuse incidence across networks (or another unit of analysis) can be the result of measurement errors or causal factors such as structural and security effort related properties of providers [97]. In this chapter, I focus on the causal factors. Figure 3.1 describes the different factors that influence abuse rates.[1] The primary cause of incidents is, of course, attacks. That relationship is moderated by two other factors: security and exposure. Neither of these factors directly cause incidents; they only influence the extent to which attacks result in incidents.

There are many definitions of security, but it generally refers to the degree in which the computing resource or service is protected against attacks so as to preserve confidentiality, integrity and availability of

1 I gratefully acknowledge the contributions of Rainer Böhme, who had the original idea for the model, and of the participants of the Dagstuhl Seminar 16461 "Assessing ICT Security Risks in Socio-Technical Systems" who helped to further articulate it.

resources. It is the opposite of vulnerability, which is one way in which it can be empirically approximated. Security, or vulnerability, can be influenced by the efforts of the defender, such as the adoption of certain controls or maturity models. It is important to separate controls and efforts from actual security. The former captures actions of the defender, the latter is the result of these actions, which may or may not be the intended or expected outcome. In many scenarios, the impact of a control on the actual security level of an organization is unknown.

The other mediating factor is the degree to which a provider, or another class of defenders being studied, is exposed to a certain threat. This is often referred to as the "exposure". Size is one example. Larger hosting providers have more customers and hence a higher probability of one of those customers being compromised. The business model can also increase exposure. Customers of cheap hosting services running popular content management systems are more likely to be compromised than professional hosting customers with their own security staff.

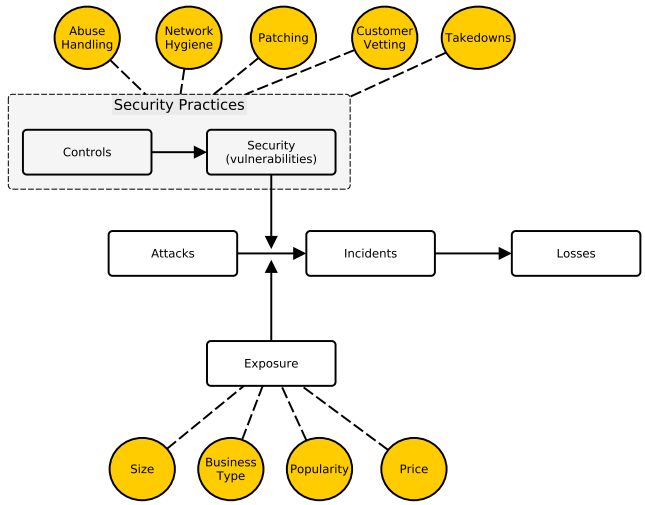


Figure 3.1: Causal model of abuse incidents linking attacks with the exposure and security practices of defenders as moderating factors that influence how attacks translate to incidents

The yellow ovals in Figure 3.1 contain examples of indicators of security and exposure. Some of them have already been found in prior work to correlate with incident rates. For security, prior studies have found that bad network hygiene and out-dated software is correlated with higher levels of abuse [73, 90, 97]. Such indicators might not clearly distinguish between controls and actual security, which is why we connect them to both, through the label of "security practices". A well-known indicator of exposure is the size of the network. Some security

metrics try to take this into account by simply dividing the number of incidents in a network by the number of IP addresses advertised by the network [85, 95, 100]. Price is another example. Cheap or free services are more prone to be abused by miscreants, leaving their providers more exposed [97].

With this causal model in hand, I can more precisely articulate the core idea of this chapter. I want to infer security performance of a provider from the abuse rate. Ideally, one would measure security independently, but often this can only be done by collecting partial indicators at best – e.g., hygiene indicators or patch levels for webstack software – or it is not possible at all. I would like to test to what extent performance can be estimated reliably as a latent factor that is driving the abuse volume.

The model illustrates that this approach assumes we can control for exposure and attacks. The former I will include in our models via a number of indicators, which I will collect for the whole population of hosting providers across the market. The latter we cannot observe directly and I will include as a random variable. In other words: I assume that attacks are randomly distributed across the attack surface. In Section 3.4, I will test how reasonable this assumption is via a simple simulation. Note that to the extent that this assumption does not hold, it will increase the size of the error term of the model, i.e., leave more variance in the abuse data unexplained and ultimately lead to greater uncertainty in estimates of the security of providers.

3.3 DATA

Abuse Data

Not all abuse feeds can be used to answer questions about security as the choice of feeds has to be balanced against the type of research question [108]. And since we are interested in the security of hosting providers, I use seven data feeds that include incidents typical for hosting services: malware-related and phishing abuse.

The malware data is provided by the Stopbadware Data Sharing Program and contains feeds from a number of volunteer companies and research institutions for the entire duration of 2015 [122]. The dataset contains URLs and IP addresses associated with and observed to spread malware. These companies use different methodologies for collection and criteria for inclusion, and furthermore the data shared by these organizations does not necessarily reflect their complete view of malware URLs. The phishing data is extracted from two sources: Anti-Phishing Working Group (APWG) [123] and Phishtank [124]. Both datasets contain IP addresses, Fully Qualified Domain Names (FQDNs) and URLs associated with phishing. Table 3.1 provides a summary of

Table 3.1: Data Feeds.

	Period	Organizations	Incidents	Abuse Type	Provider
APWG	2015	5,496	376,796	Phishing	APWG.org
Phish	2015	4,287	139,130	Phishing	Phishtank
SBW1	2015			Malware	Stopbadware DSP
SBW2	2015			Malware	Stopbadware DSP
SBW3	2015	1,580 - 7,208 **	11,976 - 376,561 **	Malware	Stopbadware DSP
SBW4	2015	(ranging between)	(ranging between)	Malware	Stopbadware DSP
SBW5	2015			Malware	Stopbadware DSP

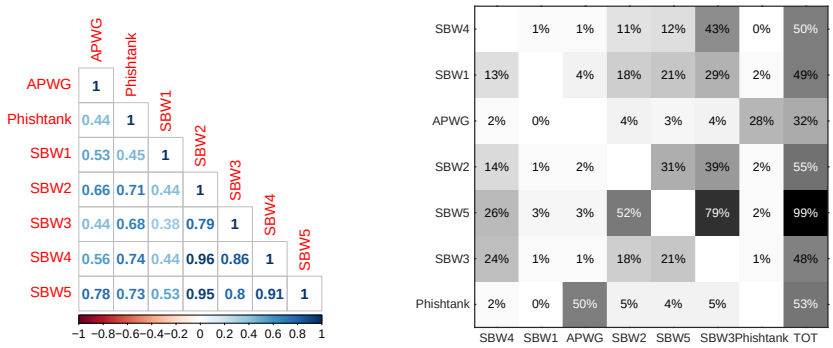
** Due to the terms of the data sharing agreement, we only report aggregated ranges for SBW data

the abuse feeds, the number of abused organizations and the number of incidents they had according to each feed over the course of 2015. Note that I employ and combine data from multiple feeds with respect to each type of abuse to ensure better results as previous work has advocated for [108].

For each dataset, I count the number of observed events per provider. Constructing such an incidence metric involves several design choices regarding the unit of analysis, attribution of incidents to the responsible units and counting the number of incidents per unit which I have more extensively discussed in the previous chapter. The metric I define as event per provider is the number of unique (2nd-level-domain, IP-Address) pairs recorded per provider in every abuse feed.

Recall that most concentration metrics choose Autonomous Systems (ASes) as the unit of analysis [93, 95, 100] and associate events with AS owners based on the BGP prefix announcements for each AS (also see Chapter 2). The AS owner, however, often merely routes the traffic for the IP address and has no administrative responsibility for it. In other co-authored work [44] however, I have developed an improved attribution approach based on WHOIS data, as it tells us to what organization an IP address is assigned. It provides a better approximation of who is responsible for abuse associated with that address than routing data can provide. The difference in using organizations rather than ASes as the unit of analysis has substantial repercussions. Some organizations operate several ASes, while in other cases several organizations may share a single AS. We found that, on average, one AS harbors seven organizations. From the total set of organizations that are found in WHOIS data, I select the hosting providers through a series of steps which I explain in Section 3.4.

Figure 3.2a provides a correlation matrix of the abuse counts across the seven feeds. The numbers underline an earlier point: abuse datasets are heterogeneous and noisy. Even sets that observe the same type



(a) (b)

Figure 3.2: Correlations between incident counts and overlap of reported 2nd-level-domains in feeds. Darker shades represent more overlap. Final column indicates percentage of feed information already contained in all other feeds combined.

of abuse, may be weakly correlated with each other. The correlation coefficient between the abuse count in Anti-Phishing Working Group (APWG) and the one in Phishtank, for example, is just 0.44. Among the malware feeds, SBW1's count also has a 0.44 correlation with the counts from SBW2 and SBW4. Figure 3.2b, alternatively illustrates the overlap between the abuse feeds in terms of what percentage of 2^{nd}-level-domains (2LDs) reported as abusive is shared among the feeds. The right most column in this figure illustrates the overlap of each feed with all other feeds combined. In other words it expresses the exclusive contribution of each feed to the overall combined data.

Hosting Data

To construct a mapping of the hosting provider market, I employ several data sources and build on techniques used in previous work [44, 100]. The mapping approach to identify hosting provider organizations is based on (i) IP ownership data from Maxmind's WHOIS API [107] and (ii) passive DNS (pDNS) data from DNSDB [111] generously provided by Farsight Security. The passive DNS data contains Fully Qualified Domain Names (FQDNs) and IP addresses that have been queried on the web and detected by Farsight's sensors in 2015.

Using the aforementioned datasets, we are able to capture several properties of hosting organizations that we can use as proxy measurements for their exposure (see Figure 3.1): (i) the total number of IP addresses allocated to an organization, (ii) the number of IP addresses allocated to the organization that are associated with domain names (i.e., those observed in passive DNS data), (iii) the total number of 2LDs

hosted by the organization, (iv) the number of IP addresses that are associated with at least 10 2LDs (a proxy for shared hosting), and (v) the number of 2LDs on shared IPs hosted by an organization.

3.4 HOSTING PROVIDER MARKET

My starting point for constructing a mapping of the hosting provider market is to map the entire IPv4 space to corresponding organizations based on the Maxmind WHOIS data. This gives us the total population of organizations to which IP addresses are allocated, as well as the number of IPs allocated to each organization. I then use pDNS data to construct the aforementioned structural organization properties discussed previously based on what has been passively observed in DNS traffic over the duration of 2015.

I define hosting providers as the subset of organizations for which we have observed at least 30 2LDs, a deliberately chosen low threshold to minimize false negatives. All other organizations are considered non-hosting organizations. Figure 3.3 illustrates the distributions of the allocated IP space to all organizations, the subset which have been observed in DNSDB and the subsets of hosting and non-hosting providers respectively.

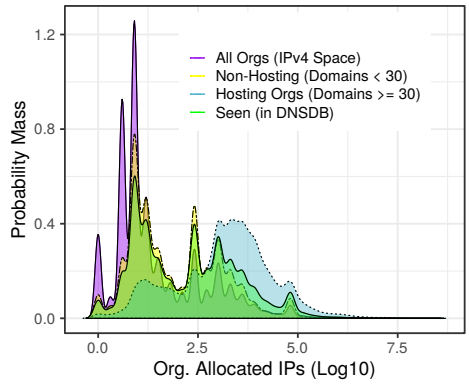


Figure 3.3: Comparing the distributions of organizations according to WHOIS IP allocation data versus the subset observed in pDNS data versus the subsets defined as hosting and non-hosting.

A comparison of the distributions of 'all' organizations and those 'observed in DNSDB' (respectively indicated by purple and green bands in the figure) demonstrates that DNSDB provides a reasonably unbiased view of all organizations, and thus of providers, as the shapes of the two distributions closely follow the same pattern, especially for organizations that are allocateds more than 10 IP addresses. This is consistent with previous research, which found that DNSDB offers a reasonably unbiased view into the entire domain name space [112].

We do however see discrepancies between the two distributions for organizations with less than 10 IP addresses. DNSDB has less visibility into this subset of small to very small networks. Given our threshold of only 30 2LDs, the probability is very low that these organizations with very few allocated IP addresses represent a significant segment

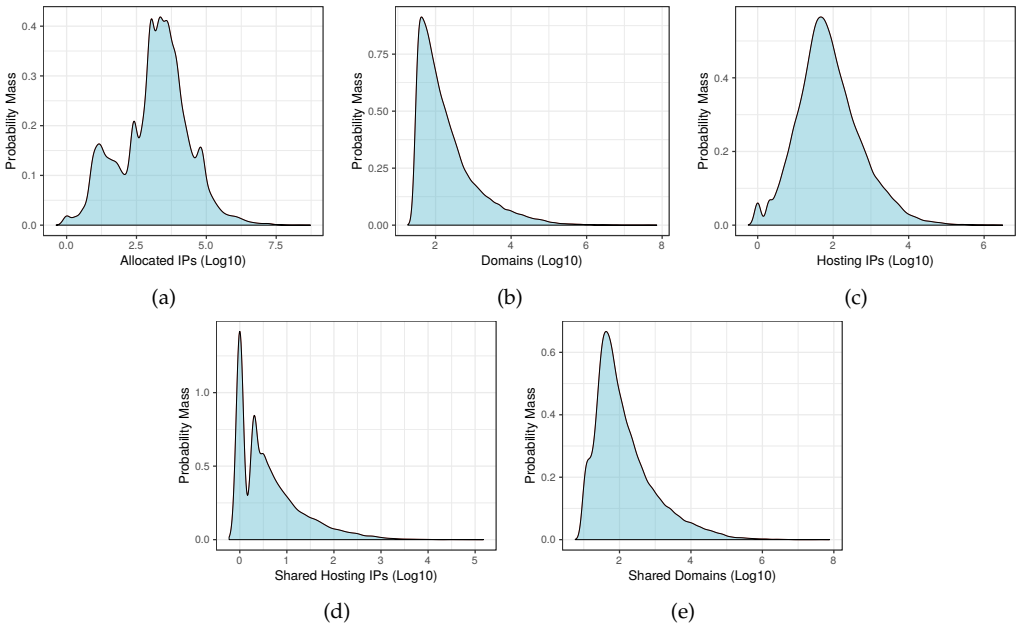


Figure 3.4: Distributions of hosting providers over various organization properties that may be used as proxies for the provider's exposure to abuse.

of the hosting provider market. Note from the distribution of hosting providers in Figure 3.3 (indicated by the blue band) that the bulk of these providers have been allocated between 300 to 10,000 IP addresses.

Given the definition of the hosting providers, we can empirically construct a picture of the aforementioned 'exposure properties' of each hosting provider by employing the passive DNS data. Figure 3.4 plots the distributions of these properties for all hosting providers, i. e. organizations that match the definition.

In terms of exposure, note that these properties not only capture size, but also include information about the business model of the provider. Three types of hosting services are related to the properties: dedicated hosting (one domain per server), shared hosting

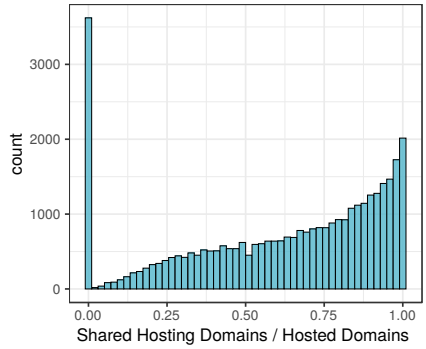


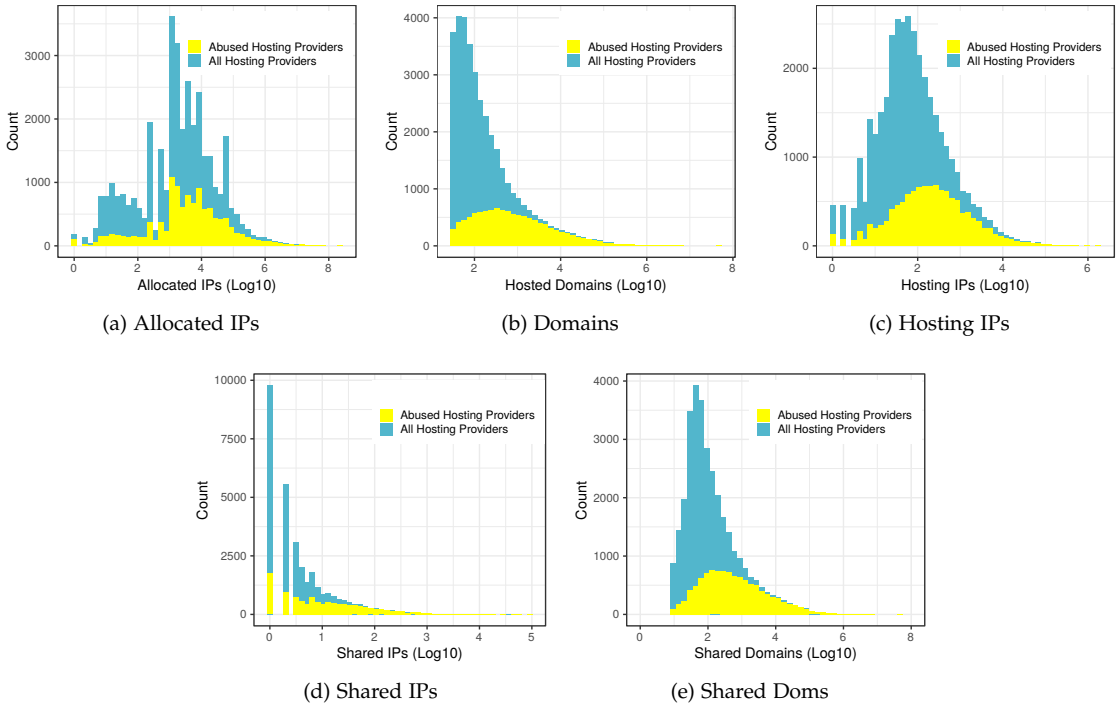
Figure 3.5: Distribution of shared hosting across the hosting market. The plotted histogram depicts the number of hosting providers with various ratios of shared to dedicated hosting across based on pDNS data

(multiple domains per server),
and services without domains (e.g., data centers or perhaps no hosting services at all).

Together, the properties capture the mix of these three types of services for each provider. Figure 3.5 illustrates the ratio of hosted domains that share the same IP address with at least 10 other domains to the total number of domains hosted by a provider as a histogram – i.e., shared hosting. The peak on the left of the figure is the population of providers with no shared hosting at all. Going from left to right, an increasing portion of the domains of a provider reside on shared hosting. In other words, the provider is increasingly dependent on shared hosting as its main business model in webhosting.

For brevity, I will not go into more detail about the provider mapping and instead refer the reader to [44] for a more in depth analysis of the market.

3.5 EXPLORING OBSERVATION BIAS IN ABUSE DATA



(a) Allocated IPs (b) Domains (c) Hosting IPs

(d) Shared IPs (e) Shared Doms

Figure 3.6: Overlaid distributions of hosting providers with observed abuse events (yellow) vs all identified hosting providers across the market (blue) over their different estimated exposure properties

I now first explore how my collected abuse data relates to the overall population of providers. The first thing that jumps out is that just 34% of all providers has at least one abuse incident in one of the seven abuse feeds. So even the combined dataset lacks observations on the majority of the market. This would be even more skewed when using only a single dataset: they cover between 5-22% of all providers.

To explore what subset of providers have abuse events, Figure 3.6 shows histograms for each of their exposure properties. Each plot shows the distribution of providers with abuse events (yellow) as an overlay on the distribution of all providers (blue). On each indicator, we see the same pattern: virtually all large hosting providers are present in the abuse data, while this ratio drops rapidly for medium-sized and small providers, where just a fraction is associated with an incident. More precisely, abuse incidents have been observed for almost 99% of the large providers (i.e., providers with 10,000 or more domain names).

One reason for this pattern is exposure: large providers have such a high exposure to attacks that the probability of incurring a single abuse incident becomes 1. That being said, there could also be observation bias at work. Perhaps the methods that generate the abuse data, whether based on automated tools or volunteer contributions, are less apt at observing incidents in smaller and medium-size networks. I test this possible explanation via a simple simulation.

Understanding the potential observation biases in the abuse data is a key consideration in constructing abuse concentration metrics as previous research points out [91, 100]. One way to identify bias is to compare the datasets against other sources of abuse data. Kührer et al. [114] for example compared abuse blacklists against each other and against data they collected themselves. In a way, I have done something similar by using seven datasets which in my case all display the same pattern.

While such comparisons are helpful, other datasets are not ground truth. They are also typically collected with similar collection methodologies. There is no ground truth for abuse data, of course. Observations are actively avoided by adversaries, and the best observation methods can at best hope to achieve a useful partial view. I therefore complement the analysis via a simulation that tests to what extent the observed pattern is consistent with a pure exposure effect. In other words, can observed patterns be explained from the attack surface of providers?

For this purpose, I assume that attackers, in the search for domains to comprise, attack the *domains* that they discover at random with a fixed probability. The specific behavior of individual attackers of course will depend on their capabilities and the types of exploits that they are able to carry out. So I am not assuming that individual attackers attack domains at random. But what I am assuming here is that the joint behavior of all attackers may be modeled as all domains having a

fixed, yet small probability of getting attacked by some attacker group. In other words, that attacks may be considered as a random variable over the range of domains.

Note that my datasets (see Table 3.1) also mainly capture cybercrime that involves domain names. Therefore, the number of domains of a provider is a useful proxy for its attack surface. If each domain has a fixed probability p of being abused, then the probability of a provider not being abused is $(1-p)^n$, where n is the total number of domains that it hosts. Conversely, the probability of a provider being abused is equal to $1-(1-p)^n$. Note that for all providers, I may obtain n from the exposure properties of the provider which I discussed earlier in the process of mapping the hosting provider market. Then, using a maximum likelihood estimator, I may estimate p from the observed abuse data which results in a value of $p = 0.0025$. Given this estimated probability, Figure 3.7 illustrates a 'separation plot' [125] of the predicted and observed abuse status of all hosting providers.

What this plot demonstrates is the degree to which the calculated probability of abuse per domain agrees with the actual observed abuse data, thus providing some confirmation that my assumption regarding the randomness of attacks may indeed be reasonable. Here, the horizontal axis, and the trend line respectively illustrate all hosting providers and the probability with which I predict they will be abused, sorted in an increasing order. A green tinted vertical thin line in the plot represents a provider for which an abuse event has been observed in the abuse feeds. More dense and darker green areas of the plot indicate a high density of providers with observed abuse events, light green or white areas indicate fewer or no such providers. The concentration of abused providers towards the right side of the plot illustrates the large degree to which the estimation results and the observed abuse data in Figure 3.6 are consistent.

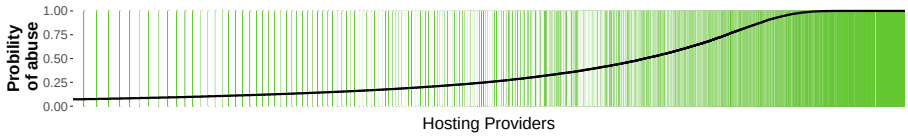
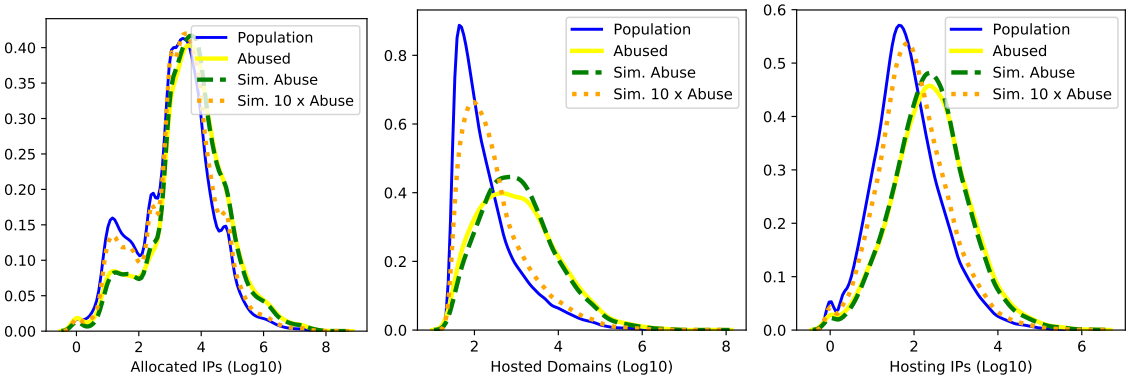


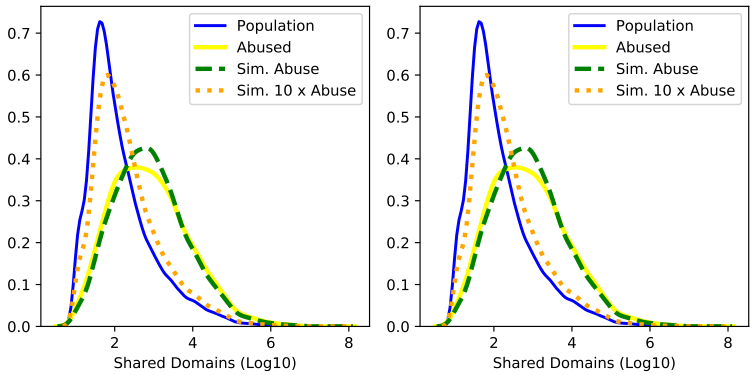
Figure 3.7: Separation plot of predicted versus observed abuse of hosting providers.

Next, I run two sets of simulations. First, I randomly select domains from the total population of domains and generate abuse incidents for the hosting providers of those domains, until we reach the same volume of incidents as I have observed in the combined empirical datasets. Next, I follow the same process, but generate 10 times more abuse incidents than the observed volume of abuse.

I then compare the distributions of providers over the different exposure properties in Figure 3.8 based on the simulation results. The first simulation, generating the same number of events as in our empirical data, produces a distributions that are highly similar to that of the empirical abuse data. The second simulation shows that as the volume of abuse increases, the distribution of abused providers approaches that of the total population of hosting providers. Another way to put this is that if we assume that all providers incur at least one abuse incident per year, which anecdotal evidence from hosting providers would suggest is not unreasonable, then the total number of incidents would be at least one order of magnitude larger than those observed by the seven abuse feeds combined. They see less than 10% of all incidents at best.



(a) Allocated IPs (b) Domains (c) Hosting IPs



(d) Shared IPs (e) Shared Doms

Figure 3.8: Non-biased distribution of abuse over population of hosting providers

These results of the simulations as well as the results depicted in the separation plot, suggest that patterns observed in Figure 3.6 are not

the result of observation bias, but rather of attacker dynamics and the random nature of the attack generation process. The simulations also provide further support for the modeling decision that I will revisit in the subsequent section, namely to model the attacks as a random variable.

Having established that there is no clear evidence for bias regarding certain providers, we can move on to the question of how to estimate security performance as a latent variable from the array of abuse datasets.

3.6 MODELING SECURITY PERFORMANCE

We are now in a position to test whether we can infer the security performance of a provider from the abuse data. Going back to our causal model (Figure 3.1), the main idea can be summarized as follows: if we are able to adequately control for exposure and we correctly assume that we can model attacks with a random variable, then the main driving factor in the abuse data is the security performance of providers. That is once other factors are taken into account the only remaining factor that drivers abuse volume should be the effectiveness of provider security efforts. When these are less effective we should expect more abuse and vice versa when the efforts are more effective. We can then try to infer this effect as a latent variable from the abuse datasets.

The simple simulation in the previous section provides support for the choice to model attacks as a random variable. The simulation was able to reproduce the empirical distribution of abuse events over the hosting market by modeling attacks as random process over the attack surface, as measured by the exposure indicators. This may be related to how attackers search for different domains to exploit, nevertheless it allows us to make simplifying yet reasonable modeling assumptions about attacker behavior. Note that the simulation results also suggests that our exposure indicators capture an important portion of the exposure factor. A more precise test was conducted by Tajalizadehkhoob et al.[97]. Using the same indicators, they were able to explain more than 80% of the variance in two phishing datasets as a function of exposure. This suggests that these indicators allow us to adequately control for exposure.

Of course, the proof of the pudding is in the eating. i will test whether my assumptions indeed hold by testing the predictive power of the estimated security performance: what portion of the remaining variance can be explained by providers' security performance, after having controlled for exposure effects. Before we get to that step, though, I first discuss the statistical approach I propose: estimating

a latent variable for each provider through a model based on Item Response Theory (IRT). What makes this approach suitable?

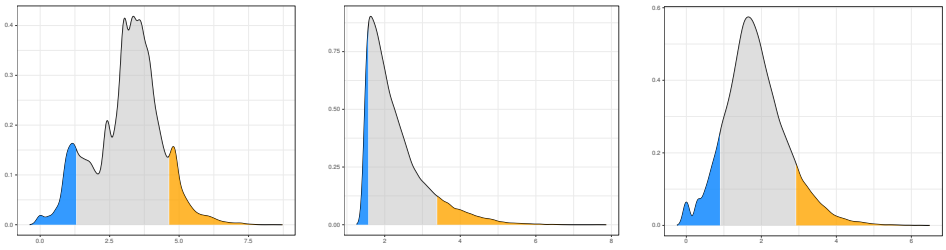
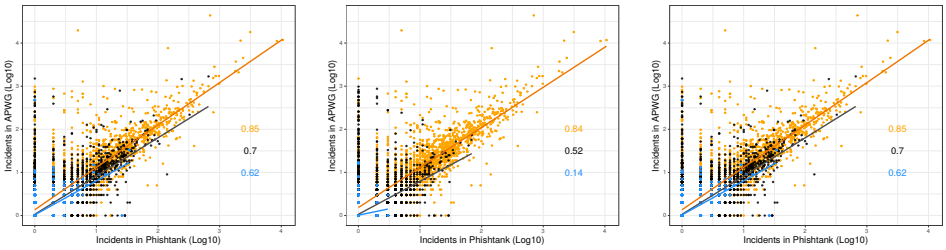
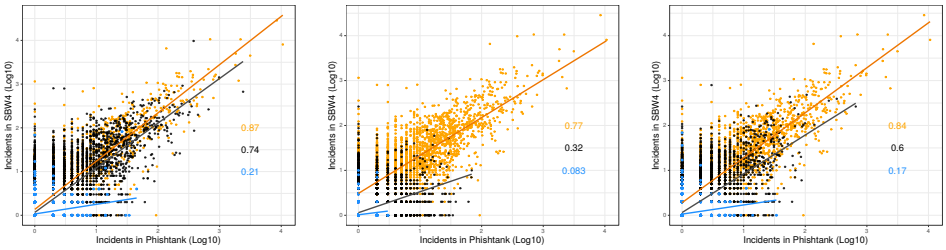
The answer lies in understanding the key requirement for this task: to estimate performance from a wide array of abuse data sources. Given the noisy and heterogeneous nature of abuse data, making reliable inferences about the security performance of providers requires us to model performance over a range of abuse data sources [108]. Earlier work however, has not provided an elegant way to aggregate information from an array of different abuse datasets. There have been two basic approaches: estimate performance separately per abuse dataset or merge all abuse data into a single set.

This first approach, estimating separate models, produces different results for different abuse types – e.g., [3, 97]. At the level of individual providers, this can generate wildly different outcomes expectedly, which is clearly undesirable for a benchmark. One solution is to average, or otherwise aggregate, benchmarks that are calculated from each individual abuse feed – e.g., in Chapter 2 I used a Borda count method. This is slightly better, but the method of aggregation introduces all kinds of artifacts into the benchmark which, again, can significantly impact the ranking of individual providers.

The second approach has been to simply merge the different datasets into a single abuse metric (e.g., [36, 85]). This means a lot of information is lost. The largest sets will drown out the signal of smaller sets, while smaller sets are not necessarily less valuable. They might capture abuse events that are harder to observe, such as the location of command-and-control servers, but very relevant to the overall abuse landscape. The merging might also average out differences in the susceptibility of providers for certain types of abuse, but not for others. Any performance benchmark would benefit from taking that into account.

Table 3.2 highlights some of the complex, yet meaningful, relationships among abuse data sets. It compares abuse data from three of our abuse feeds in relation to some of the exposure properties of the providers. I compare two data sources capturing the same type of abuse and two data sources capturing different types of abuse. The first comparison, using phishing data from Phishtank and APWG, contains signals about measurement errors. Some providers have a high incident count in one feed, but a low count in the other feed. As both feeds capture the same type of abuse, we suspect this difference is mostly due to measurement error. This demonstrates how (in)consistently the abuse data captures this particular type of abuse. The second comparison, between Phishtank and the SBW4data, also shows inconsistencies for providers. In addition to measurement errors, this also signals differences in the susceptibility of the provider's infrastructure to different types of abuse. Clearly the consistency of the strength and reliability of signals varies depending on which part of the hosting provider

Table 3.2: Exposure properties of abused organizations in relation to various abuse types for the 10th, 10-90 and 90th percentile of the providers (respectively indicated by light blue, gray and orange colors).

Properties	**Owned IPs**	**Hosted Domains**	**Hosting IPs**
Empirical Distribution			
APWG / Phish			
SBW4/ *Phish*			

Note: Reported numbers in plots represent correlation strength

population we inspect as indicated by how strongly the different data points for different segments of the population correlate.

To meaningfully capture the different signals within the abuse data and to overcome the aforementioned issues, we apply techniques from Item Response theory [126, 127] to our abuse data. In the subsequent sections, we explain the general approach, specify the model, estimate the latent variable of security performance and then test its predictive power against independent abuse data.

3.7 IRT MODEL SPECIFICATION

To better capture the information in each of our abuse feeds we specify an analytical model which draws from Item Response Theory (IRT). Applications of IRT models have previously been explored in risk assessment [128]. However, IRT models are most commonly used to measure the effects of an unobservable latent capability of a student – let's say math skills – from how well (s)he performs in a range of tests. The student examination metaphor can provide a good intuition of how our approach works. We approach incident numbers in each of the abuse feeds as an indicator of how good or bad each student performed in an exam consisting of several questions, with questions corresponding to our abuse feeds. Needless to say, hosting providers are the equivalent of students in this metaphor. Just as exam questions vary in terms of subject and difficulty, we assume that our various abuse feeds reflect similar properties. Some abuse events are more difficult to detect than others, which is reflected by the number of incidents observed per provider in different abuse feeds. Also note that exam questions often have overlap in terms of their subject matter, and we consider our 2 phishing and 5 malware feeds to reflect a similar property as our analogy.

The model is graphically illustrated in Figure 3.9. For every abuse feed $j = 1, ..., k$ and for every provider $n = 1, ..., N$, the abuse incident count Y_{nj} follows a Poisson distribution

$$Y_{nj} \sim \text{Poisson}(\lambda_{nj})$$

with

$$\ln(\lambda_{nj}) = \ln(\text{E}(Y_{nj}|\theta_n, \mathbf{x}_n)) = \gamma_j + \mathbf{x}_n^T \boldsymbol{\beta} - \alpha_j \theta_n. \tag{3.1}$$

This model consists of k Poisson regression models, one for each abuse feed $j = 1, ..., k$, where γ_j is a feed-level intercept, $\mathbf{x}_n$ is a vector of exposure-related covariates for provider n with coefficient vector β (shared across feeds), and θ_n is a continuous latent variable that captures structural variation in abuse counts across providers. This latent variable has an additive effect on every abuse count, but the sensitivity of each abuse count to the latent variable, α_j, is different for every abuse feed. We constrain $\alpha_j > 0$, $j = 1, ..., k$, so that a higher value for the latent variable θ_n leads to lower expected abuse counts for every feed. As such, a higher positive value for the latent variable θ_n represents more effective security performance, and a negative value represents less effective security performance. Hence, the latent variables θ_n quantify the level of effectiveness of the security practices of each provider. The feed-level sensitivity parameters α_j represents the difficulty of mit-

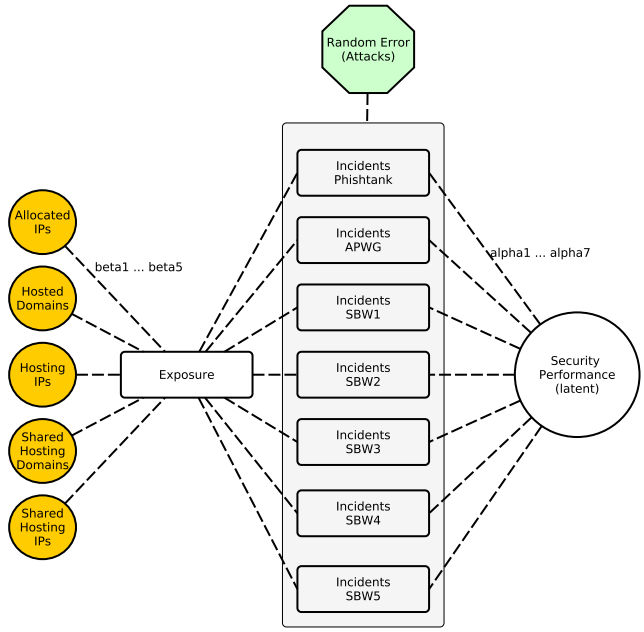


Figure 3.9: IRT model jointly explaining variation in incident counts for all abuse feeds. The model simultaneously relates the number of incidents empirically observed in each abuse feed with the exposure properties of providers, attacker behavior which may be seen as random error given the exposure of providers, and finally the security performance of providers as the main factors driving abuse volume.

igating the abuse measured by each feed $j = 1, ..., k$. We further specify θ_n as draws from a standard normal distribution

$$\theta_n \sim \mathrm{N}(0, 1).$$

The variance of the latent variable distribution is constrained to 1 for identifiability, since all the sensitivity parameters α_j are freely estimated.

Intuitively, this model disentangles the portion of the variation in incident counts that is due to varying levels of exposure, and attributes the remaining variation to varying levels of security performance of the providers, after considering what part of the variation is random noise from attacker behavior after having accounted for exposure.

3.8 ESTIMATION RESULTS

To infer the security performance of providers from abuse data, we input the incident numbers from all abuse feeds into the IRT model and estimate the parameters of the model (see Equation 3.1) using MCMC simulation. The model uses the exposure related variables (numbers of

hosted domains, shared hosting domains, allocated IPs, hosting IPs and shared hosting IPs) to control for exposure related effects. Note that some of our exposure related variables capture the attack surface while others the business type of providers. The model uses a logarithmic transformation of the independent (exposure) variables as input. Part of the variation in incident numbers that cannot be attributed to exposure make up the values for our latent security performance variable.

I performed full Bayesian inference of the model parameters and the latent variables by means of Markov Chain Monte Carlo (MCMC) sampling [129]. I used weakly-informative prior distributions for this purpose

$$\gamma_j \sim \mathrm{N}(0, 10), \qquad \ln(\alpha_j) \sim \mathrm{N}(.5, 1), \qquad \boldsymbol{\beta} \sim \mathrm{N}(0, 3)$$

reflecting a relative ignorance of their true values. I carried out MCMC sampling using Stan [130] with the `rstan` R interface. I ran 4 chains for 1500 iterations each, with 750 warmup samples. This resulted in a total of 3000 MCMC samples.

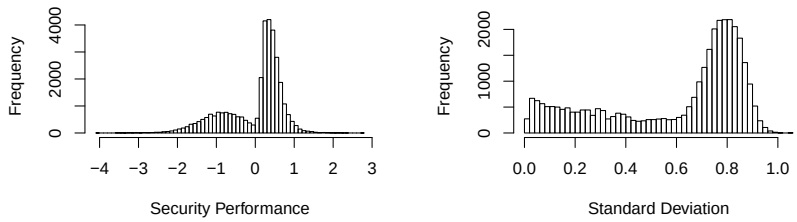
Table 3.3: IRT model parameter values for all abuse feeds

	Parameter for	Mean	SE-Mean	SD	2.5%	97.5%
γ [1]	APWG	-7.13	0.01	0.03	-7.20	-7.06
γ [2]	Phishtank	-6.09	0.00	0.03	-6.15	-6.04
γ [3]	SBW1	-9.06	0.00	0.04	-9.13	-8.98
γ [4]	SBW3	-5.10	0.00	0.03	-5.15	-5.04
γ [5]	SBW4	-5.09	0.00	0.03	-5.14	-5.04
γ [6]	SBW2	-5.72	0.00	0.03	-5.77	-5.66
γ [7]	SBW5	-6.27	0.00	0.03	-6.33	-6.22
β[1]	Owned IPs	-0.75	0.00	0.01	-0.76	-0.73
β[2]	Hosting IPs	-0.36	0.00	0.01	-0.39	-0.34
β[3]	Hosted Domains	3.82	0.00	0.03	3.76	3.88
β[4]	Shared IPs	1.25	0.00	0.01	1.22	1.27
β[5]	Shared Domains	-1.96	0.00	0.03	-2.02	-1.91
α[1]	APWG	3.19	0.01	0.03	3.14	3.25
α[2]	Phishtank	1.83	0.00	0.02	1.80	1.87
α[3]	SBW1	2.50	0.01	0.03	2.45	2.55
α[4]	SBW3	2.14	0.00	0.02	2.10	2.17
α[5]	SBW4	1.80	0.00	0.02	1.77	1.83
α[6]	SBW2	2.13	0.00	0.02	2.09	2.17
α[7]	SBW5	2.31	0.00	0.02	2.27	2.35

The MCMC algorithm converges towards the parameter values summarized in Table 3.3 with $\hat{R}$ values close to 1, which indicate convergence of the sampling algorithm [131]. The table reports the estimated

posterior mean value of each parameter along with the 95% credible interval in which we estimate the value to be.

The first set of parameters, γ, are intercept values that set the baseline of abuse levels for each abuse feed. The second set of parameters, β, capture the effect of each exposure variable on the incident numbers in all abuse feeds. The third set of parameters, α, capture how much the security performance of providers affect abuse levels in each of the feeds. Intuitively this is similar to the difficulty of exam questions from our analogy of the IRT approach. For example $\alpha[5]$ which has the lowest value among the α parameters, tells us that the security performance of providers has the least effect on lowering incident numbers within that feed. By analogy, it is a hard question to get right on a student exam.



(a) Security performance of providers (b) S.D. of performance point estimates

Figure 3.10: Distributions of security performance (latent variable)

The final model parameter, our latent variable θ, represents the security performance of providers, which is what we are interested in. Based on our modeling results, Figure 3.10 illustrates the distributions of the posterior mean of the latent variable and the posterior standard deviations for all providers respectively. As stated earlier, security performance is measured on a continuous scale where larger positive number represent more effective security performance and negative numbers represents less effective performance. Notably, Figure 3.10b demonstrates that the posterior standard deviation of a considerable portion of the measured performance levels is large. The large posterior standard deviation simply quantifies our own lack of certainty about the true value of the latent variable. For that subset of measurements our confidence in estimated values is low. We explain why this large standard deviation occurs shortly here after.

Next, Figure 3.11 illustrates the distributions of the estimated latent security performances of all hosting providers across the market. It depicts the posterior mean of the latent variable by black dots along side the 95% credible interval of the latent variable values as colored error bars. An orange color bar indicates providers for which abuse

has been observed while gray colors indicate providers for which no abuse has been observed according to our abuse feeds. Of course and as stated before, the security performance values are reported along a latent variable scale where positive numbers represent better security performance and negative numbers vice versa. To gain an intuition for the reported security performance values, note that the last term in the IRT model specification from Equation 3.1, which models the effect of security performance on abuse volume, is characterized by a subtraction of some value from the overall right hand side of the equation. This basically means that a positive value for security performance results is modeled as having a decreasing effect on the volume of incidents for each provider.

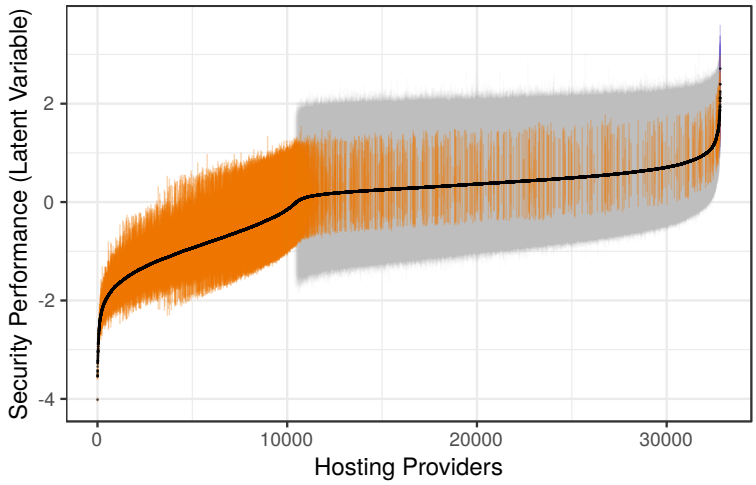


Figure 3.11: Estimated provider security performance with 95% credible interval band. Security performance is reported on a latent variable scale along the y-axis with positive numbers representing better security performance and negative numbers vice versa. All hosting providers identified across the market, are represented along the x-axis in improving performance order.

Going back to the matter of the certainty of the estimates, roughly stated, the posterior standard deviations for two thirds of the providers is larger than 0.5. The remaining security performance values have a standard deviation smaller than 0.5 and capture the level of security performance with more certainty.

The consequential larger credible intervals are the result of a large range of potential θ values being plausible, given the observed abuse data and our model. Improving on this requires larger samples and more abuse data with a stronger signal to noise ratio. This is a limitation of our model and of the data. A second factor that leads to large credible intervals for a subset of the providers is that these coincide

with providers for which zero abuse incidents have been observed, which also happen to be small hosting providers. For such providers it is difficult to disentangle whether the lack of abuse incidents is due to their small exposure or due to their high security performance. Therefore our model is not able to accurately capture how well they perform in terms of their security.

Another reason for larger credible intervals is when incident counts in different abuse feeds are wildly different, combining high and low abuse rates. As Table 3.2 illustrates, for a certain selection of providers, abuse feeds show very different incident counts. Further analysis of these cases however, revealed that all have a posterior standard deviation smaller than 0.5 associated with their estimated security performance.

Despite the uncertainty about the exact values of the latent variable, in the next section we will see that taking the posterior means as a simple point estimate of security performance proves to be quite robust and can be used to generate good out-of-sample predictions.

3.9 ROBUSTNESS AND PREDICTIVE POWER

Given the estimated security performance levels, we can examine how much of the variation in incident counts can be explained by the mean point estimate of the latent variable. To do so, I construct a GLM model of the incident counts which includes the latent variable as an explanatory factor, in addition to the exposure related factors. As we did in our IRT model calculations, the incident counts are fitted to a Poisson distribution of the same form as described in Equation 3.1 ⊛. To test the predictive power of the approach, I measure the security performance value repeatedly, each time leaving out one of the abuse feeds. I then use the measured security performance to explain the variance in incident counts in the independent abuse feed that was left out of the estimation process. This way, we may cross-validate the results and can examine the predictive power of the calculated security performance values. Table 3.4 shows how different models for the SBW1 dataset explain the observed variation in incident counts, where security performance was measured from the other abuse datasets.

⊛ *Note the log linear modeling, e. g. using the Poisson distribution, is both typically a statistically appropriate choice for fitting count data, but also in this case observed to provide a reasonable fit to the observed incident counts based on residual analysis and goodness-of-fit indicators in the Bayesian fitting process used to estimate the latent variable*

Model (1) is a baseline model which only includes a constant baseline value as an explanatory factor. Model (2) adds the number of hosted domains as an explanatory factor, and model (3) includes all exposure-related indicators. Model (4) is the final model and adds security performance as an explanatory factor.

As indicated by the log-likelihood, AIC and dispersion of the models, model 4 is a considerable improvement over the models with only exposure effects. In addition, the pseudo-R^2 values presented in the table indicate that exposure alone explains 78% of the variation in

abuse counts, while latent security practices add an additional 20% to the explained variance – or 91% of the variance that remained after controlling for exposure.

Table 3.4: Poisson GLM regression with *ln* link function

	Dependent variable:			
	SBW1 Incident Counts			
	(1)	(2)	(3)	(4)
Hosted Domains (Log_{10})		1.96^{***}	-1.58^{***}	1.70^{***}
		(0.01)	(0.11)	(0.09)
Hosted Shared Domains (Log_{10})			1.98^{***}	-0.53^{***}
			(0.10)	(0.08)
Allocated IPs (Log_{10})			0.43^{***}	0.05^{***}
			(0.02)	(0.02)
Hosting IPs (Log_{10})			0.26^{***}	0.09^{***}
			(0.03)	(0.03)
Shared Hosting IPs (Log_{10})			1.42^{***}	1.22^{***}
			(0.03)	(0.03)
Security Performance				-1.84^{***}
(latent variable)				(0.01)
Constant	-1.01^{***}	-7.95^{***}	-7.24^{***}	-9.15^{***}
	(0.01)	(0.04)	(0.06)	(0.07)
Observations	32,822	32,822	32,822	32,822
Log Likelihood	−63,479.84	−21,701.49	−15,556.22	−3,142.87
Akaike Inf. Crit.	126,961.70	43,406.97	31,124.44	6,299.74
Dispersion	422.43	9.71	12.39	0.12
Pseudo-R^2	0.00	0.68	0.78	0.98
Note:				*p<0.1; **p<0.05; ***p<0.01

The coefficients for the explanatory variables in the model can be interpreted as follows. I use model (4) as an example, needless to say that the other models can be interpreted in a similar fashion. Lets take the coefficient value of 1.70 for the number of hosted domains as our primary example. This value indicates that, while holding all other independent variables constant around their mean, increasing the number of hosted domains by 1 unit (the equivalent of multiplying the number by 10 due to the $\log_{10}$ scale of the variable) results in the expected number of incidents of the provider being multiplied by $e^{1.70} = 5.47$.

The interpretation is slightly different for the coefficient of the security performance variable: −1.84. Here, the GLM model suggests that increasing the variable by 1 unit, while holding all other variables constant, reduces the number of incidents of the provider by a factor of

$e^{-1.84} = 0.158$ – in other words, by 84%. Increasing the latent variable by 1 basically means increasing security performance by one standard deviation. The range from -2 to 2 includes 95% of all providers.

The inverted coefficient signs of the number of hosted domains and shared hosting domains between model (2), model (3) and model (4) are due to the interactions between exposure variables, as some of them are derivative of others. For instance, 'large hosted domain size' results in 'large hosted shared domain size' as well. Modeling each of the exposure variables separately shows a positive significant effect for each one, which is in line with what we observe in model (4).

We have repeated the same procedure for all abuse feeds. That is, we measured security performance based on all feeds except one and then explained the variance of incidents counts in the feed that was left out. The main results are summarized in Table 3.5. The total explained variance in the incident numbers, using both exposure and security performance, ranges from 75% to 99% of the total variation. The key finding here is that the security performance variable reliably adds to the explained variation of each individual feed that has been left out of the security performance estimation process. This suggests that the variable is able to capture the security performance of providers reliably enough to have considerable predictive power. The coefficient value for the latent variable in these models ranges between -2.13 to -1.54 and consistently shows a significant relation with the incident counts as the dependent variable.

Table 3.5

	Incident Counts According to Abuse Feed						
Variance Explained by	SBW1	SBW2	SBW3	SBW4	SBW5	APWG	Phishtank
	Relative to intercept only baseline model						
Exposure	0.78	0.86	0.83	0.89	0.85	0.70	0.83
Exposure + Security Performance **	0.98	0.95	0.99	0.96	0.96	0.75	0.89
	Relative to exposure only model						
Security Performance **	0.93	0.64	0.98	0.65	0.79	0.19	0.40

Note: ** Calculated from all feeds excluding feed indicated by column heading

The additional explained variance for all results, indicated in the bottom row of Table 3.5, is remarkably high for such noisy data on such a multicausal phenomenon.

Two feeds stand out however. The predictions for the APWG feed and the Phishtank feed are less strong than those for the malware-related feeds. We speculate that this might be due to an imbalance in the number of feeds that have been used as input to the IRT model for calculating the security performance variable. In both instances, the modeling procedure involves the use of five malware-related abuse feeds, leaving only one additional phishing feed that has been used to measure the security performance. Therefore the security performance variable calculations are slightly skewed towards values dictated by the malware related feeds. In future work, this seeming lack of unidimensionality can be further explored by estimating a two-dimensional item-response model, in which the security performance of providers is allowed to vary along two dimensions. Presumably, one of these dimensions will be more strongly correlated with phishing abuse, and the other with malware abuse. Such an analysis may reveal to what extent these different types of abuse feeds can be seen as measurements of the same latent trait, and as a consequence, how sensitive our security performance estimates are to the selection of abuse feeds used to estimate them. In addition, finding a diverse set of abuse feeds with minimal redundancy will likely improve the robustness of the estimated security performance.

3.10 RELATED WORK

Many studies use abuse feeds as their primary source of data on security incidents, with different objectives.

A few studies have looked at abuse patterns across single or multiple threats, with the intent to explore or explain what factors correlate with abuse levels. The main implication of these studies is that concentrations of abuse are the result of poor security practices. Zhang et al. found that network hygiene – measured by the normalized number of misconfigured systems – is correlated with a range of abuse incidents as observed by various blacklists [85]. Liu et al. have also been able to predict data breaches using various indicators of network hygiene and a combination of security incident data feeds [73] . The underlying logic is that security hygiene practices of providers drive abuse rates across different threats. Or reverse: that one could infer effective security practices from combining different abuse data sources. Note that this study merged all of their abuse data into one combined data set, which might mean that the largest set overwhelms all others and thus the study finds a relation with that specific set of observations of abuse. A similar approach, but then at the organization level, was conducted by Edwards et al. They assessed the security performance of organizations from externally collected indicators of their security posture, and find that it correlates to abuse data [132]. Shue et al. also utilize abuse

information from multiple abuse sources and combined them into a single set to examine the connectivity characteristics of networks with unusually high concentration of blacklisted IP addresses [92]. Vasek et al. combine abuse data sources to identify risk factors for webserver compromise [90]. Soska and Christin extend this work, by developing predictive systems to identify whether individual websites may be used for malicious purposes in the near future through the extraction of features from website contents, and verifying their predictions using a combination of abuse data feeds [133]. Our work is similar to this body of work by following the logic that is behind correlating indicators with abuse. However, our work mainly differs in its unit of analysis, namely hosting provider organizations, and how we utilize our abuse datasets towards our goal of inferring security performance.

A separate body of work has looked at concentration of abuse events in certain networks [36], Internet Service Providers (ISPs) [80, 116], Autonomous Systems [93, 100, 105], countries [3, 117, 118], organizations [73], payment providers [13], registrars [5], registries [112], and other agents. The idea is that such concentrations are amenable to intervention. They are interpreted to reveal attacker economics – such as scale advantages – or defender economics – such as a lack of security investment by some agents because the cost of incidents is externalized to others [91, 97]. This chapter contributes to this body work by offering a systematic explanation for abuse concentrations, replacing speculative interpretations of what they imply about security efforts or attacker preferences. This work is most closely related to [97] in which Tajalizadehkhoob et al. propose analytical models to explain abuse concentrations based on exposure. I build upon this work with a different modeling approach, based on IRT. I also build upon [44] to construct a mapping of the hosting provider market and explore the issue of bias in my abuse data.

Others studies have experimented with mixing abuse data to infer reputation scores for individual hosts or IP addresses to help protect services. One such approach is the idea of threat exchanges. Thomas et al. examined the usefulness and limitations of mixing multiple sources of abuse information for this purpose [25]. This work helps illuminate the relationships among abuse data sources, or the lack thereof, but their analysis has very different purpose and does not provide any insight into the security efforts of larger aggregates, such as networks or providers.

Orthogonal to the subject matter of this chapter is the wide range of problems associated with incident and abuse data, on which a lot of security research is based. In Chapter 2 for instance, I systematically walk through some of the difficulties associated with creating operator benchmarks based on multiple data sources [100]. Clayton et al. highlight considerations that need to be made before intervening

based on abuse concentration metrics [91] an important part of which is measurement bias and possible artifacts that it produces. Kührer et al. attempt to quantify the measurement bias of a combined set of malware blacklists in comparison to independent data sources [114] and find its effects to be considerable. These studies combine data sources to reduce the effects of bias, use independent datasets to examine consistency and use multiple measurements to indicate stability of results over time. The implication being that there is minimal/negligible effects from bias. Pitsillidis et al. reflect on the various spam abuse data collection techniques and the variations in abuse data that can produce different findings [108]. The authors investigate the noisy nature of spam abuse data and how it is necessary to combine multiple data sources to more appropriately investigate security questions. Metcalf and Spring compare the contents of 25 different blocklists and surprisingly find very little overlap between the contents of the blacklists [109]. This work takes such issues into account and carefully explores the bias in our various data sources to ensure minimal effects on its results.

3.11 DISCUSSION AND CONCLUSIONS

The success of many industry and government initiatives to combat cybercrime relies on the ability to empirically track the security efforts and progress of various market players. Abuse data is a critical resource in that endeavor, but also a rather unruly one. This chapter addressed the question of whether one can infer a reliable measurement of security performance of hosting providers from an array of different abuse feeds.

Abuse datasets are notoriously noisy, highly heterogeneous, incomplete, biased and driven by multiple causal factors that are difficult to disentangle. Earlier research has managed to address some of these issues, but here I have presented a more comprehensive approach that takes all of them into account. I have applied this approach to the hosting sector, which is associated with a large portion of all observed abuse events.

I have also presented a causal model for the generation of abuse data that is implicitly behind much of the discussed empirical research. Furthermore, I have undertaken an exploration into observation bias, which showed that its impact is limited in terms of the distribution across the hosting market. The heart of the proposed approach is a modeling approach based on Item Response Theory (IRT), which estimates the security performance of hosting providers as a latent variable from an array of abuse datasets. The Bayesian nature of the approach also means that we can quantify the (un)certainty that we have about the security performance signal, as the security performance of each provider is expressed as a distribution. The proof of the pudding is in the eating, of course. I have tested the robustness of the approach via

out-of-sample predictions. And I found that the security performance measurements can predict a large amount of the variance in abuse incident counts, after controlling for exposure. In short, our results demonstrate that a careful modeling of abuse data can generate robust and reliable signals about the security performance of providers.

There are also limitations to this approach of course. Due to the noisy nature of the abuse data and the limitations of the model, the certainty in our security performance factor for providers can be low, for a significant part of the hosting provider market, most notably the smaller providers. That being said, the fact that the modeling approach is able to quantify uncertainty is in itself an improvement over existing approaches. Notwithstanding the uncertainty, the results turned out to be remarkably robust and powerful, as shown by the out-of-sample predictions. Prediction power for the two phishing datasets was lower. One answer would be to select a more balanced set of datasets. A less arbitrary approach would be to experiment with two-dimensional latent trait models, which I intend to undertake in future work.

In sum, I would argue that the current approach can help improve the security incentives by reducing information asymmetry in markets where abuse incident can be observed and associated with defenders. It provides a basis to measure the impact of security controls and practices on performance, thus providing a more empirical basis for industry security practices and government oversight.

4

EVALUATING HOSTING PROVIDER REACTIVE REMEDIATION EFFORTS

In Chapter 2, I systematically walked through some of the challenges of developing metrics for and comparing hosting provider security postures on their basis. Subsequently, in Chapter 3 I discussed a comprehensive modeling approach to develop such metrics based on observing how frequently abuse incidents occur at each provider. These developed metrics are reflective of ***proactive*** *security efforts as they signal how well providers prevent security incidents from happening in the first place. They capture the idea that if providers proactively secure their infrastructure through effective security practices, they should have to deal with less incidents, something which should also be reflected in lower numbers of and less frequent incidents.*

But in Chapter 2, I also discussed a second type of metric for comparing hosting provider security postures which is based on how timely incidents are remediated by the providers once they have already occurred. In this chapter, I systematically examine how abuse data may be employed to construct such metrics and compare hosting provider security performance on their basis. These alternative metrics, which capture incident remediation times, reflect the ***reactive*** *security of providers and yield a more direct measurement of security effort. They capture the idea that once incidents occur, independent of how exposed or secure the providers' infrastructures are, more secure providers should remediate incidents faster. As such, they have the added benefit of their outcomes not being affected by provider exposure.*

*I first present an empirical analysis of how the actions of defenders and attackers impact the time required to remediate abuse involving websites. Upon examining two leading industry data sources of abuse data, I then identify two fundamental challenges that must be overcome in order to develop this second type of metric and to draw valid inferences about provider security on their basis: (****i****) measurement errors and (****ii****) multi-causality. I identify different causal factors affecting remediation time, notably the behavior of attackers and the shared responsibility of different defenders. I then disentangle and quantify the causal impacts by constructing survival regression models with a set of explanatory variables that capture each actor's behavior and subsequently derive metrics for comparing hosting provider security postures.*

This chapter is based on my third peer-reviewed study on the subject of metrics development [102] (still to be published), focusing on the development of security metrics that are reflective of ***reactive*** *provider security efforts.*

4.1 INTRODUCTION AND BACKGROUND

"Abuse" data, for example phishing, malware spreading, or botnet command and control (C&C) domains, are circulated industry wide and used in the day-to-day fight against cybercrime. Many industry sources collect, track, and share data on website abuse to protect customers and remediate compromise. This brings immense benefits such as scale, operational visibility, and impact to name a few. Moreover, abuse data are invaluable to researchers, potentially shedding light on the effectiveness of defender responses. To understand which defenders and their respective responses are more effective at remediating abuse however, we first need to be able to compare defenders.

As such, this chapter focuses on how to leverage abuse data towards comparing defender responses from data on their incident remediation times: i.e. the time required by each to clean up a compromised domain or neutralize a maliciously established one. To this end, I collect and analyze data from two leading industry sources, `Google Safe Browsing` [17] and `Spamhaus DBL` [134] which track the abuse of domain based resources hosted within various networks. While the primary purpose of this data is to protect users by identifying malicious activity, their sources also indirectly track the time required to remediate abuse incidents.

Internet intermediaries, which are often in a key position to respond to abuse, are broadly acknowledged to play an important role in defending against and the remediation of abuse [43, 51, 78, 100, 101, 112]. Driven by the fact that our collected abuse data largely captures incidents in hosting networks, my aim is to compare the remediation efforts of hosting providers using this data. Achieving this, constitutes a key step towards understanding which hosting providers are more effective at defending and perhaps understanding why.

While most prior studies characterize and compare hosting provider security efforts based on incident counts [97, 101], i.e. abuse concentration, this study measures and compares responses more directly by constructing metrics based on the amount of time taken for abuse incidents on hosting provider networks to be remediated. The study also differs from prior work that measure remediation times as it does not focus on the effect of specific countermeasures (c.f. [19, 135]) or the effect of interventions (c.f. [77, 136]) on remediation times but rather on hosting providers as defenders themselves and the effect they have on remediation.

Despite our data from the industry sources having been used in numerous prior studies, it should come as no surprise that real-world measurements of remediation time are prone to all kinds of errors and artifacts that affect observations which are implicitly or explicitly discussed as limitations in prior work. The way in which most industry

abuse data are collected is underspecified, often intentionally, to combat reverse engineering by attackers and competitors. This leads to an opaque view of the data generation process and its measurement errors which only become transparent once data has been analyzed in-depth.

It is also important to realize that incident remediation times are the resulting outcome of the entangled actions of multiple actors [43, 53], most notably webmasters, intermediaries such as hosting providers, the third parties who discovered the abuse and, last but not least, attackers.

Notwithstanding this jumble of entangled measurement errors and causal factors, many studies directly attribute observed variation in remediation times to specific actors - for example webmasters [26, 135, 137] or network operators [36] - sometimes overlooking the other driving factors. Others attribute variation to the introduction of specific countermeasures - for example notifying some of the involved parties in incidents and in a position to act upon the notifications [19, 54]. The studies that deploy a randomized experimental design in studying their factors of interest (e.g. [19, 54]) can make such attributions with some confidence but are also typically the ones that are focused on measuring the effect of a specific countermeasure. In observational studies like this study however, it is important to rule out that differences in remediation time are not caused by factors other than the factor of interest, i.e. the defenders that are being compared, and to control for the causal effects of other actors.

The challenge therefore, is to identify the influential casual factors in, and where possible overcome the measurement errors of, remediation time data towards drawing robust inferences about how various hosting providers compare in terms of their remediation efforts. This study should also help anyone using third party abuse data, to avoid its potential pitfalls, by demonstrating an approach of how such data can be more safely consumed and interpreted.

A necessary first step towards this goal then, is to systematically identify factors that drive variation in empirical observations of remediation times. I therefore develop an analytical model in Section 4.2 that identifies sources of variance – in addition to webmaster behavior – based on the data generation process underlying prior work and careful analysis of the data sources.

Given the widely used industry abuse feeds, a detailed description of the data and my methodology for extracting remediation times from each dataset are described in Section 4.3.

I then analyze the data in depth in Section 4.4, while Section 4.4.1 illustrates how differences between when a cleaned resource is dropped from the blocklist contribute to significant measurement errors. Section Section 4.4.2 demonstrates how widely used metrics in the literature and inferences drawn therefrom may be affected by these measurement errors.

To draw, inferences about the causal drivers of remediation efforts, I subsequently construct several explanatory models of the remediation time data in Section 4.5. Using a set of proxy variables pertaining to each agent's behavior described in Section 4.5.1, I then investigate how the multiple agents identified by the analytical model influence remediation time by disentangling their causal effects. While prior work has focused more on webmaster efforts, in Section 4.5.2, I find that hosting provider efforts and attacker behavior each have a strong impact on remediation times. For the DBL data, provider and attacker-related factors together explain nearly 60% of the total variation in remediation times. For GSB data, I again find that provider characteristics also explain 12% of the total variation.

To deal with the measurement errors that I identify in the datasets, I triangulate my findings by comparing DBL malware data to Safe Browsing data in Section 4.5.3 in order to reach more robust inferences about and comparisons of hosting providers. While some of the effects that I find are consistent and therefore may generalize, I also find that certain provider-level explanatory variables have inconsistent effects (i.e., the same characteristic is associated with longer remediation times in one dataset and shorter times in the other). This both demonstrates and reinforces the need for more rigorous triangulation in future studies of a similar nature. I then briefly reflect on this study's empirical findings in Section 4.5.4.

Next, Section 4.6 summarizes the related work, highlighting the similarities and differences, and how this study extends prior work.

I finally conclude by providing recommendations in Section 4.7 and articulating the main lessons learned from the study applicable in future research.

The main contributions may be summarized as follows:

- I provide an analytical model describing sources of variation in remediation time measurements. It generalizes existing models from studies which are largely based on measuring incident counts (c.f. [97]), to the case of tracking incidents over time.
- I unpack how measurement errors manifest in remediation time measurements and how they threaten the validity of inferences drawn therefrom.
- I demonstrate a modeling approach to compare defender responses and draw robust causal inferences from noisy remediation time measurements. As such I extend prior work by explicitly modeling and controlling for the behavior of a more complete set of actors.
- I demonstrate that hosting provider and attacker behavior significantly affect remediation times.

- I demonstrate an approach of "triangulation" across multiple data sources to limit the impact of remediation time measurement errors on causal inferences drawn from the data.
- I discuss and propose alternative solutions for improving data collection and observational research based on remediation time measurements.

4.2 DATA GENERATION MODEL FOR REMEDIATION TIMES

As a first step, I have identified a broad set of factors that influence the data generation process and observations of remediation times by analyzing prior research. I incorporate these factors into an analytical model, which captures the sources of variance in remediation time observations (Figure 4.1). The model separates influential factors into two main sources of variance: measurement errors and causal factors. An intuitive way to think about these factors is that they explain why two empirical observations of remediation time may differ.

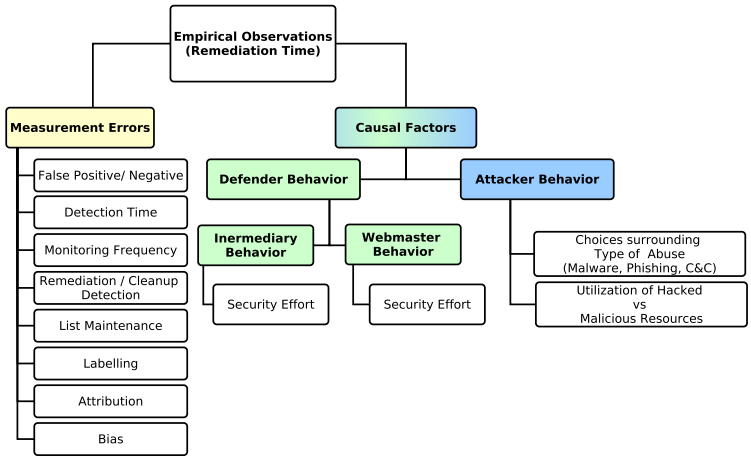


Figure 4.1: Analytical model of variation sources in remediation times

I first briefly explain the various factors of this analytical model. At suitable points I also refer to literature in which some of the elements have been explicitly or implicitly discussed as findings or limitations.

I start with discussing variance in observations driven by measurement error and how the factors listed under this category in Figure 4.1 have been discussed in the literature.

Prior work discusses several types of measurement errors that may affect remediation time data. For example, imperfect vulnerability and abuse detection instruments lead to false positives and negatives (cf [25, 55, 135, 138, 139, 140]). Similarly, imperfections can also lead to

incorrect detection of remediation (cf [77, 135]). When remediation is carried out by making the harmful resource unavailable, researchers may incorrectly deem malicious resources as clean that are in fact temporarily unavailable due to networking problems [28].

Perhaps to avoid or lessen the impact of such imperfections, some researchers rely on real-time abuse data compiled by industry to measure remediation time [135, 141]. Such abuse feeds have certain advantages, including scale, impact and visibility, as they are often supported by more resources and expertise than academic measurements can bring to bear. Furthermore, because being placed on an industry blocklist can have immediate negative consequences, webmasters are incentivized to take action and drop their blocked resources off the list. The choice of consuming industry data however, comes at the expense of loss of control and accepting an opaque view of the data generation process, along with its potential measurement artifacts and errors.

Additional errors arise when datasets are augmented with labels for further analysis, such as assigning responsibility for incidents to specific entities [97, 100] or differentiating the type of abuse or vulnerability. For example, prior work discusses that certain types of abuse are more difficult to detect. Attackers may also actively try to evade detection [21, 28, 43, 141, 142]. Attacks that are difficult to detect may not show up at all in remediation data or may appear shorter in duration due to difficulties in detecting their initial onset.

Measurement frequencies also lead to imprecision over exactly when incidents and, later, remediation occur [60, 135, 141]. This can in turn bias remediation time measurements [3, 44]. For example, all such measurements systematically underestimate the remediation time of incidents that have not been resolved by the end of a study or ones that existed at the start of a study. As such, remediation time measurements at best represent estimates of true remediation time. As a result, researchers regularly employ survival analysis (e.g., [19, 21, 141]), a statistical technique designed and well equipped to deal with remediation time estimations and its inherent biases. Other forms of biases may also arise, for example when measurement instruments are employed at different rates during a study or in relation to specific data subjects [18, 135].

Having enumerated many of the factors leading to measurement errors, I next turn to discussing the relevant causal factors driving variation in empirically observed remediation times. These relate to attacker and defender actions, that is, the efforts of those perpetrating and remediating abuse. I first discuss factors relating to attackers.

Prior work has demonstrated that the way in which attackers abuse resources plays a role in how well they protect them against detection and remediation. For example, Moore and Clayton demonstrated that remediation times were slower for phishing attacks perpetrated by

criminals using botnets to host resources, compared to those compromising websites [60]. Similarly, in their investigation of search result poisoning and search redirection attacks Leontiadis et al. have demonstrated different remediation times for abused resources, depending on whether the resources are owned or more closely guarded by the attackers themselves or not [20]. More specifically, Leontiadis et al. demonstrate that traffic brokers employed to redirect traffic for instance to fake pharmaceutical websites, which are under the full control of attackers, are remediated at a much slower rate than the actual pharamaceutical websites themselves, or the compromised websites that are used to advertise the latter in order to increase their search rankings.

Some prior work also discusses the fact that attackers may use a single resource for several purposes at different times or abuse a resource again if not properly remedied [100, 135, 141]. Multiple infections, e.g., one domain name being repeatedly compromised can prolong remediation time. Finally, whether attackers misuse hacked resources or maliciously create their own, also plays a role in when or whether potential remediation will take place at all [19, 135]. Attackers may also use bullet-proof hosting services that notoriously ignore abuse complaints and help miscreants dodge remediation [100, 143]. It is unfortunately impractical to expect attackers to assist in remediating their own maliciously created resources. Consequently, the burden of remediation of maliciously-hosted content falls entirely on defenders who do not directly control the resources in question.

On the defender side, the literature discusses the role of actors who influence remediation efforts. I categorize these into two main types: webmasters and intermediaries ⊛. Webmasters understandably play a crucial role in remediating abuse on their systems. They typically have the most control over a resource and its security. They are also most directly affected if their traffic plummets as a result of appearing on a blacklist. Much of the research has focused on the efforts of webmasters in combating abuse [26, 30, 135].

⊛ *For a comprehensive discussion on the roles of various actors in combating cybercrime, I refer the reader to [43, 53].*

Yet there is reason to believe that when it comes to remediating abuse, intermediaries play at least as big of a role as the webmasters. Intermediaries include hosting providers, network operators and registrars. They are often better positioned to observe abuse directly and process third party abuse reports. In some contexts (e.g., shared hosting [78]), providers are capable of directly implementing remediation. Researchers have investigated the importance of hosting providers [28, 44, 51, 141, 144], ISPs [80], and registrars [4, 112] in remediating abuse.

Of course independent third parties such as governments, law enforcement, and organizations harmed by or tracking abuse can also influence remediation. Prior research has established that the incentives of the party requesting remediation often influence whether and how fast abuse is remediated [67]. More indirectly, third parties create cer-

tain institutional and socio-economic environments in which criminal behavior may be more or less persistent [143, 145]. Such third party influence however, is indirect and acts through the medium of the efforts of defenders and attackers discussed earlier.

Defenders jointly influence how timely or effective remediation is. Webmasters impact remediation through patching and cleaning of their abused resources, intermediaries through their proactive and reactive security efforts and third parties through monitoring, notifying and information sharing about vulnerabilities and abuse. Finally, remediation times also depend in large part on the efforts of attackers in designing attacks that evade detection.

4.3 INDUSTRY ABUSE DATA

For this study I collect data from two industry leading abuse feeds to gather observations on the remediation time of abuse incidents. Both feeds are operational real-time blocklists actively used to prevent access to online resources or warn Internet users of potential harm that they may face through accessing the flagged resources. I first provide a short description of each data source and its data generation process to the extent that is publicly known in Section 4.3.1 and subsequently explain the methodology for extracting observations on hosting provider remediation responses from these data sources in Section 4.3.2.

Data Feeds and Collection Methodology

Here, I collect data from the Google Safe Browsing [17] and Spamhaus DBL [134] abuse feeds. These jointly capture information about the remediation of three distinct abuse types, namely: malware, phishing and botnet Command-and-Control (C&C) infrastructure. To provide a more detailed overview, I split the data collected from these feeds into distinct datasets each focusing on the remediation of a single type of abuse. A summary of this data is provided in Table 4.1.

Table 4.1: Overview of abuse feeds and data extracted from the feeds over the period of 2017-07-17 / 2017-10-31

Dataset	Organizations	Hosters	URLs	2LDs	Abuse Type	Abuse Feed
GSB	3,618	2728	295,326	127,833	Malware	Google Safe Browsing [17]
DBLM	765	675	-	3,684	Malware	Spamhaus DBL [134]
DBLP	1,678	1,357	-	94,403	Phishing	Spamhaus DBL [134]
DBLCC	1,351	1,135	-	36,980	C&Cs	Spamhaus DBL [134]

The GSB dataset, derived from Google's Safe Browsing feed, is comprised of URLs that attempt to automatically download malware onto

victim computers when visited. While the genesis of this feed can be traced to a published research paper [138], the feed today is essentially a black box for which only specific details about its collection and scanning methods are publicly known. What is known is that Google regularly flags URLs which point to harmful content, notifies appropriate parties (webmasters / domain owners) of the flagged domains and monitors cleanup progress through rescanning flagged resources. This feed is used by Google to displays warnings to Internet users if flagged domains appear in search results, or when users navigate to such domains within the Chrome, Firefox or Safari browsers. If a domain owner signs up, they will be notified via Google's webmaster tool and can take action if and when security issues are discovered. Domain owners, perhaps with the assistance of their hosting providers, must remove the harmful content in order to drop off the list. After harmful content has been removed, webmasters can request a rescan through Google's webmaster tool or through StopBadware's appeals process. Without a rescan request, domains are automatically rescanned 14 days after having first been flagged. The frequency by which flagged resources are automatically rescanned beyond the 14 day period is not publicly known, although Google's public pages indicate that some networks may be scanned more often than others. Google claims a negligible false positive rate [135]. Through collaboration with StopBadware, I have an hourly view of this feed and collect snapshots of its data. I refer the reader to [17, 135, 146] for more information regarding this data.

The `DBL{M|P|CC}` family of datasets are derived from the Spamhaus `DBL` (Domain Block List) feed, which is a real-time blocklist of 2^{nd}-level-domains (2LDs) with low reputation. Spamhaus deems these domains to be involved in sending, hosting or origination of spam, phishing websites, distribution of malware or C&C infrastructure of botnets. The feed is primarily intended to be used by mail server software to identify, tag or block incoming email containing domains that Spamhaus flags based on the contents of message bodies. The DBL can be used to reject spam email, as well as emails containing or pointing to harmful content in their message body. Documentation suggests that spam traps are used primarily to populate the DBL. DBL also claims that their feed has almost no false positives [134]. To achieve that, Spamhaus drops blacklisted entries if a certain expiry time has been reached. Entries are tagged with additional information on the type of abuse. These tags fall into two main categories of `abused legit` and `malicious` domains. Within each category at least 4 subcategories are identified namely, `spam`, `phishing`, `malware` and `botnet C&C`. The `DBL{M|P|CC}` datasets have been constructed based on these tags. Because my focus is on websites hosting malicious content, I discard the entries flagged for spam which typically point to end user machines rather specific hosted servers. I

have been collecting a snapshot of the entire DBL list every 15 minutes during the period indicated in Table 4.1. Spamhaus documentation suggests that flagged abusive domains are regularly monitored for remediation although it is not clear how frequent rescanning occurs. More detailed information about the DBL data feed can be found on the DBL documentation pages [134].

Definitions and Data Processing Methodology

I extract incident remediation time observations from the data, by processing the snapshots collected from each data source. Note that the GSB dataset snapshots record harmful URLs, whereas the DBL datasets record harmful domains as explained in the previous section. I defined and tracked incidents across snapshots at a domain ownership level in order to unify the unit of analysis across all datasets. That is, incidents were defined as pertaining to 2^{nd}-level-domains (2LDs), e.g. `example.tld`. In certain cases where domains may be owned at sub-levels, for example with a structure like `*.example.tld`, incidents were defined as pertaining to the sub-level domains ⊛. I employed Mozilla's public suffix list [147] to map all snapshot entires onto their defined 2LD unit of analysis.

⊛ *Such cases typically occur on cloud and shared hosting platforms or for example under TLDs like* `.co.uk`.

Given this particular unit of analysis, the remediation time of an incident was defined as the continuous timespan between its discovery, i.e. a 2LD's entrance on the blocklist, to the moment it was explicitly flagged as resolved or removed from the data.

Note that some domains have had repeated incidents and reappeared as flagged entries in the data after having been removed. In this respect, I treated non-overlapping abuse incidents that involve the same 2LD as separate incident cases of abuse, i.e. reinfection, whereas duration-overlapping incidents were merged and treated as a single incident. Figure 4.2 graphically illustrates how remediation times have been derived and reinfections treated.

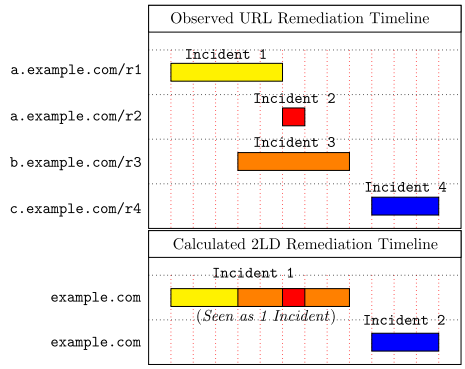


Figure 4.2: Derivation of remediation time by collapsing URL remediation timeline

I have also regularly resolved the flagged entries of the datasets, to their corresponding IP addresses during the collection of the data. I have used this information to attribute each incident to a specific

hosting provider organization by looking up the organization to which the corresponding IP address was allocated at an incident's onset, based on WHOIS information. Here, as in Chapter 3, I have made use of historical IP WHOIS data provided by MaxMind [107]. As such, I identified the hosting providers that had been in a position to take remediation action on each incident.

Note that I excluded a small proportion of organizations that were identified by the described attribution process from analysis. Their proportions may be observed in Table 4.1 by comparing the columns labeled as "`Organizations`" and "`Hosters`". First, several governmental and educational organizations were identified from specific keywords in their WHOIS data organization names and excluded from further analysis as they were clearly identifiable as not being hosting providers that publicly provide hosting services. Next, I adopted the heuristic definition of hosting providers developed in [44, 101] which I discussed previously in Chapter 3, to further exclude some organizations deemed to be too small to constitute hosting providers. As such, only organizations that hosted at least 30 2LDs were deemed hosting providers. I calculated the number of 2LDs hosted by all organizations in similar fashion to the previous chapters data using passive DNS data from DNSDB [111] collected during 2017. The heuristic threshold was then applied to exclude non-hosting provider organizations other than the previously identified governmental and educations institutions. Note that the particular adopted 2LD threshold was chosen based on a *receiver operating characteristic* (ROC) curve analysis of manually constructed ground truth data and demonstrated to have reasonable accuracy in the referenced prior work. The potential attribution errors of this process are also more extensively discussed in [97]. For more details, I refer the reader to this related work.

Given this methodology, I identified the hosting provider organizations captured in the data, tracked all of their incident remediation times and excluded non-hosting organizations and their observed incidents from further analysis.

4.4 EXAMINING REMEDIATION DATA

My use of `GSB` and `DBL` data as industry sourced data, means that I have a less than complete and rather opaque view of the data generation process, potential measurement artifacts, and the biases of each data source. Therefore, the important questions to address first, are concerning the extent to which my derived incident remediation times may be considered accurate, and what types of potential measurement errors they are affected by. A lot here depends on how incidents have been tracked and how frequently they have been monitored, which are factors that are outside of my control. Therefore, my first step is

take a closer look at the derived remediation data through the lens of the analytical model (discussed in Section 4.2) in order to identify potential measurement errors and biases. To this end, in Section 4.4.1 I first examine the data using survival analysis, comparing data sources and abuse types. I then split the data by hosting providers in Section 4.4.2, highlighting the challenges of drawing comparisons among hosting providers based on the derived remediation times to understand how and to what extent the measurement errors in the data affect comparisons of hosting providers.

Measurement Errors

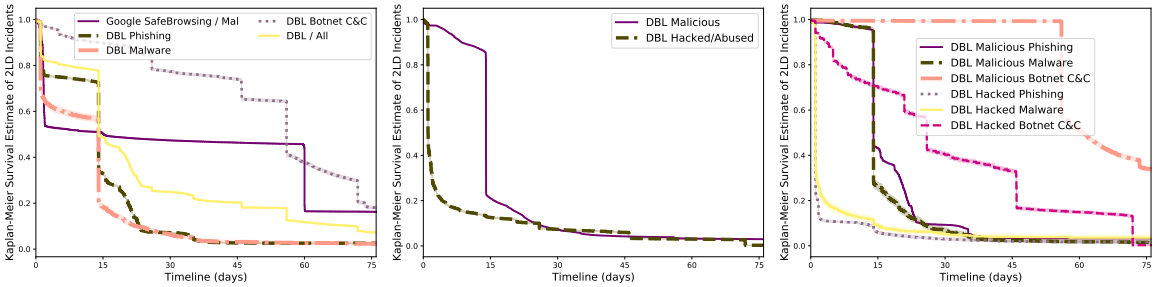
To analyze the derived remediation times I use survival analysis as a tool to inspect my data in-depth. In addition to survival analysis being a well-established statistical technique to analyze elapsed time measurements, I first highlight several additional reasons for why survival analysis was chosen a as suitable tool.

First, survival analysis is a well suited technique to analyze time-interval measurements, especially when they contain censored data. Due to the limited time window of my study, many of the observations capture the onset of specific incidents but data collection does not carry on long enough to capture their remediation. These are referred to as (right-) censored data points and contain lower bound information on remediation times. For example, an incident that has *not* been remediated by the end of a 60 day study (censored), reflects the fact that its remediation time is *at least* 60 days. Note that, censored data are typically excluded from analysis in prior studies that measured remediation times. Survival analysis allows me to retain information carried by such data points by giving them less weight without excluding them from analysis.

Second, It is important to note that virtually all remediation time measurements are estimates. The more frequently monitoring occurs, the closer the estimates are to their true values. Since the monitoring frequencies of my original data sources are not fully known, it is important to treat the derived remediation time measurements as estimates. As such, survival analysis also allows me to analyze remediation times along with associated statistical error bounds. Errors may be smaller particularly when multiple observations with respect to a particular hosting provider or particular type of incident are available. In other words, by employing survival analysis techniques, I am forced to take confidence bounds into account when drawing comparisons and inferences from the data.

As such, the first step in analyzing the data, is to empirically derive the probability that incidents will *not* be remediated within a certain time by applying the non-parametric Kaplan-Meier estimator [148] to

the remediation time observations. The KM-estimator, constitutes a useful metric of comparing remediation times and additionally allows us to unearth hidden measurement artifacts in the collected data as I shall demonstrate shortly hereafter. Its results may be intuitively interpreted as characterizing the fraction of incidents that have (or have not) been remediated by a certain time since their onset.



(a) All Datasets (b) Based on DBL domain classification (c) Based on DBL abuse type

Figure 4.3: Incident remediation trends in collected datasets from GSB and DBL. Plots report the KM survival probability estimate of incidents over a timeline measured in amount of time required to remediate incidents

Figure 4.3a depicts the survival probability of incidents captured by each of the datasets plotted against a timeline relative to their discovery. Right censored data points in the datasets are observable by the illustrated survival curves not approaching the *zero* mark on the y-axis. For the GSB and DBL feeds, I respectively observe 18% and 10% of incidents not resolved by the 75th day since their discovery. Note that, had I excluded these censored data points, the results and comparisons would have been misleading and would appear to suggest that all incidents were remediated within the study period.

We can clearly see measurement artifacts in the resulting survival curves that are indicative of some of the discussed measurement errors in Section 4.2. For example, some of the observed artifacts relate to the timing and frequency of monitoring incidents. Note however, that these artifacts are to be expected as evident from the public documentation that I found on these data feeds.

Data from GSB for example exhibits a small drop on the 14-day mark (corresponding to the first time that Google automatically rescans all discovered incidents), a rather flat curve between the 14th and 60th day followed by a large drop in survival probability on the 60-day mark. Similarly, data from the DBL exhibit drops in probability over several points on the survival timeline, most prominently on the 14-day mark. Patterns observed in GSB data are indicative of irregular monitoring of incidents, especially beyond the 14th day mark. Similarly, further

analysis of DBL data (see Figure 4.3b and Figure 4.3c) also suggest irregular monitoring of incidents specifically related to those labeled as "malicious" 2LDs by Spamhaus.

The most direct comparison that can be drawn between the datasets is between GSB and malware-related incidents in the DBL which pertain to the same type of abuse. The corresponding survival curves plotted in Figure 4.3a both see a quick drop on the first initial days, but diverge thereafter. This suggest that differences across these two datasets may at least in part be driven by particular biases of each data source. This however, is not necessarily problematic as the datasets may be capturing responses to different incidents.

My analysis of the observations in each dataset show that there is indeed very little overlap between GSB and DBL in terms of the individual incidents that each dataset captures. Prior work already demonstrates that abuse feeds typically have very little overlap in terms of the incidents they capture [109] as I have also alluded to in previous chapters.

Overall however, my datasets capture the behavior of a shared set of 2,097 hosting providers, and provide me with repeated remediation time observations on incidents that occurred on their networks. In terms of the more directly comparable *malware abuse* subsets, data on a shared set of 475 hosting providers has been captured which I later use to triangulate comparisons of defenders across these two different data sources.

Comparing Provider Efforts

To compare the remediation responses of hosting providers, I start by dividing the data based on the providers to which incidents are attributed. Comparing remediation times across providers, enables us to evaluate if some defender efforts have been more effective than others. Outcomes of comparisons however, depend on how comparisons are drawn and which metrics are constructed as I demonstrate next.

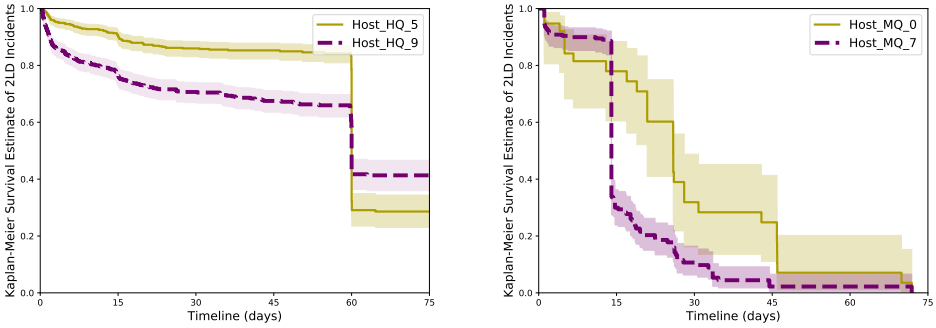
We typically observe two types of metrics constructed on top of remediation data in the literature when comparing remediation times across groups: point estimates and distributions.

Examples of point estimates include: the proportion of remediation at certain points of time in various notification campaigns compared in [54], remediation speeds of DDoS attack amplifiers between countries and between Regional Internet Registries (RIRs) compared in [136], and remediation times of botnet C&C domains hosted in certain countries [145]. Other work considers distributions. We see for example comparisons of Internet intermediaries in terms of the statistical distribution of their reaction times to C&C domains hosted on their servers in [3, 145],

or comparisons of groups of data subjects based on the cumulative distribution of their reaction times in [135].

Prior work has regularly drawn inferences from such metrics, implying that they reliably capture differences among data subjects. However, making this assumption may not be justifiable in the presence of the measurement errors that I found in the data. Therefore, I first investigate how the measurement errors of my data reflect in several commonly constructed metrics, what pitfalls they may lead to, and which types of metrics are less impacted by measurement artifacts, i.e. more reliability characterize differences among providers.

To this end, I sample 30 providers from the data and construct metrics for this set of providers. I limit the sample to providers where at least 50 incidents were observed. I also segment the providers into the lower 10th (`LQ`), 10-90 (`MQ`) and 90th percentile (`HQ`) based on their size and randomly select 10 providers from each population segment to construct metrics. I then construct several metrics to compare these providers and investigate how accurately they characterize the differences among the providers based on several illustrative examples.



(a) GSB (b) DBL

Figure 4.4: Remediation trends with 95% confidence bands for two sets of providers in different size percentiles.

The first example depicted in Figure 4.4 plots the remediation efforts of some large and medium sized providers selected from the GSB and DBL feeds. Here, we can observe that measurement errors likely prompting a drop in survival probability affecting the resulting view of defenders' remediation efforts to different degrees. In Figure 4.4a, at the 60-day mark we observe a sudden drop of approximately 55 percentage points in survival probability for one hosting provider, and 25 percentage points for the other.

Despite this artifact, there appears to be a significant difference in the distributions, but these are masked if one only examines the median value: in both cases 50% of the incidents have been remediated by 60 days.

The opposite outcome happens when comparing defender efforts in Figure 4.4b. Here, the measurement error manifests in one provider's outcome only, skewing the median calculations for only this provider. Since half of incidents are supposedly remediated by 14 days by this provider, this creates a shorter median cleanup time for than the provider without such measurement error.

As such we may conclude that metrics that compare point estimates of provider remediation efforts are likely to result in misleading comparisons based on the data I have collected.

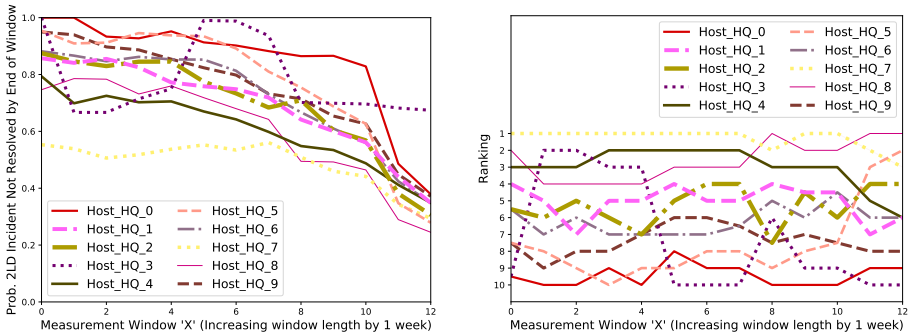
We may alternatively construct metrics and compare the providers depicted in Figure 4.4, based on their remediation speed or proportion of remediation at specific points of time. It is clear though that inferences drawn from such metrics are sensitive to the specific points in time which comparisons are made. Comparing remediation efforts at 14 days and 75 days in Figure 4.4a for example, would yield completely opposing results. In the former case they would suggest that the provider labeled as `Host_HQ_9` was more responsive while the opposite conclusions is drawn in the latter case. This demonstrates, that comparisons based on remediation speed and proportion of remediation, which are point estimates again, may also be misleading.

To avoid such misleading inferences, I conclude that metrics should ideally preserve relative differences among providers and retain information on variation along dimensions in which comparisons are made. These are properties that survival curves based on the KM estimator meet best.

Other common metrics, for example rankings do not meet these criteria and are also impacted by the measurement artifacts.

I constructed rankings of defender remediation efforts based on the proportion of incidents remediated over increasingly large time windows of my data. For GSB data I observed the autocorrelation of provider rankings before and after the 60th day artifact to drop from a Pearson-r value of 1 to 0.78 signaling a notable impact on how rankings immediately before and after the artifact relate to each other. For DBL data, the Pearson-r value approximately held a constant value of 1 before and after the 14th day artifact. During this study's period, of the 187 and 1,016 providers in GSB and DBL data that had at least 50 incidents, 177 and 140 providers respectively changed in ranking position after the observed measurement artifacts.

Rankings have additional limitations as well. Constructing them would again require comparisons of point estimates, for example the proportion of remediation at time 'X', which I already illustrated to be sensitive to errors. They also mask relative differences among providers by projecting them onto fixed ranking positions which do not preserve relative distances. It is not possible to infer that a better ranking provider has for example remediated 'Y' times as many incidents rela-



(a) (b)

Figure 4.5: Rankings of defender efforts based on remediation proportion achieved over growing time windows of captured remediation data

tive to a lower ranking provider by comparing their rankings let alone why. Even when rankings are constructed over time, they are subject to the natural variability of data over time and are impacted by measurement artifacts. Figure 4.5 which plots the constructed rankings of the 10 randomly selected providers from the HQ percentile of providers illustrates these issues.

In short, different metrics over the same data can lead to drastically different conclusions about defender remediation efforts. Given that measurement errors exist, an alternative solution may be to treat and analyze the data as a binary signal of remediation, i.e. remediated vs not-remediated incident. This would essentially require that I ignore the time dimension of my data and would be antithetical to my stated aim of evaluating security efforts based on the time required to remediate incidents. Moreover, applying such a reduction on the data, places us in the realm of studies that compare remediation efforts through abuse concentration.

At the same time it is important to realize that even in ideal scenarios, constructing metrics over my data to compare providers, does not eliminate the fact observed differences in metric values, even when comparing survival curves, may be driven by other causal factors than just the providers.

To compare provider remediation efforts, there are other options however. One could move beyond comparison of point estimates, or reduce data to a binary signal, towards causal analysis of provider differences captured by entire survival curves as outlined in the next section.

Table 4.2: Proxy Indicators used to model captured remediation data

Abbr.	Description - (Source)	Trans.	Indicator for	[min - max / mean]
RNG	*Nr. of IP addresses allocated to provider* (WHOIS)	Log_{10}	Provider Behavior	[0.3 - 8.34 / 4.98]
DOMs	*Nr. of domains hosted by provider* (*DNSDB* [111])	Log_{10}	Provider Behavior	[1.49 - 7.76 / 6.08]
IPs	*Nr. of IP addresses observed to be hosting domains* (*DNSDB*)	Log_{10}	Provider Behavior	[0.3 - 6.73 / 4.21]
SDOMs	*Nr. of domains sharing an IP with minimum 10 others* (*DNSDB*)	Log_{10}	Provider Behavior	[0 - 7.76 / 6.01]
SIPs	*Nr. of IP addresses hosting at least 10 domains* (*DNSDB*)	Log_{10}	Provider Behavior	[0 - 5.28 / 3.36]
CAR	*Nr. of Alexa top-1M ranked domains hosted* * (DNSDB/Alexa)	Log_{10}	Provider Behavior	[0 - 7.61 / 4.29]
MAR	*Median rank of Alexa ranked domains hosted* * (DNSDB/Alexa)	Log_{10}	Provider Behavior	[0 - 6 / 5.83]
SPM	Security Performance Metric ([101], see Chapter 3)	-	Provider Behavior	[-3.54 - 2.18 / -0.81]
AR	*Alexa top-1M ranking of domain with Incident* (Alexa)	Log_{10}	Webmaster Behavior	[0 - 6 / 0.03]
ML	*Malicious domain (versus hacked)* ** (Spamhaus)	-	Attacker Behavior	$\{0, 1\}$ / 0.71
AT	*Type of Abuse* ** (Spamhaus)	-	Attacker Behavior	$\{0_{cc}, 2_{ph}\}$ / 1.33

* Alexa ranks have been reversed. Highest ranked website has ranking 10^6

** Only available for DBL data

4.5 DRAWING CAUSAL INFERENCES

With a few notable exceptions (e.g., [30, 145]), studies commonly interpret differences in remediation times as the result of the actions of one particular actor [28, 144, 44, 112, 141], often the webmaster (c. f. [26, 137, 135]). While a randomized experimental design can often isolate the causal effects of particular actors, they are not always feasible. Many studies (c. f. [141, 135]), including mine presented in this chapter, are based on observational data and lack randomized experimental controls. In these studies, it is very difficult to evaluate whether it is justified to attribute all variation in remediation times to one specific class of defender. A lot depends on specific properties of the data and methodological assumptions to see if the impact of other defenders, as well as the attackers, are indeed controlled for. I argue that a more systematic and transparent approach is to explicitly model the causal impacts of the different agents (defenders and attackers), as I have done in Section 4.2. I next apply the data to the model in order to draw conclusions about what factors affect remediations time observation and how the different providers captured by the data compare in terms of their efforts.

Causal Model and Proxy Indicators

I explicitly model the impact of defenders and attackers based on the causal model illustrated in Figure 4.6. Differences in the behavior of these agents, i.e., attackers, hosting providers, and webmasters from my analytical model (Section 4.2), are hypothesized to be the primary driver of variation and the cause of variations in remediation times. For

modeling purposes I construct several proxy indicators to approximate their behaviors.

I retrospectively collect the indicators on a best-effort basis, meaning I collect pertinent additional information for abuse incidents at the time which they occurred. Notwithstanding their limitations (more on this later), I will demonstrate that even this incomplete and imperfect set of indicators has significant explanatory power. More importantly, I will show just how substantial the impacts of different agents can be. The indicators used in my statistical models of the data are summarized in Table 4.2, which also lists summary statistics and an explanation of which agents' behavior they are associated with.

As I am primarily interested in hosting providers, I control for differences in webmaster behavior through the proxy of a domain's popularity (AR), similar to methods used in [30, 104].

I also control for differences in attacker behavior observed by the proxy of type of abuse (AT) and whether attackers utilized hacked resources or maliciously created their own domains (ML). Data on the latter indicator is only available for the DBL data provided as part of the abuse event data by Spamhaus.

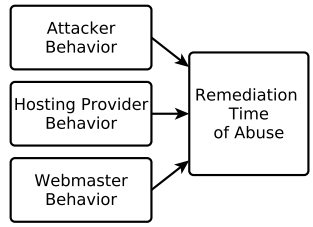


Figure 4.6: Assumed causal model linking variation in remediation times to attacker, provider, and webmaster behavior

I then utilize a number of provider level indicators to capture how differences in provider security efforts relate to their size (RNG, DOMs, IPs), business model (SDOMs, SIPs), popularity (CAR, MAR), and their proactive efforts (SPM) similar to methods used in [101, 78]. The latter indicator is essentially the result of the metrics that I produced previously in Chapter 3 captures how effectively providers prevent abuse of their infrastructure. It may serve as a signal of whether a provider's network has high concentration of abuse incidents as would be the case with bullet-proof hosting for example.

Causal Analysis

I employ Cox proportional hazards regression (CoxPH) to model and estimate the impact of the collected set of indicators on remediation times in the DBL data (and GSB data later in §Section 4.5.3). CoxPH models quantify the magnitude and direction of the effect of variables in terms of regression coefficients. They are commonly used in prior work [30, 145, 141] for datasets involving observations of elapsed time where some values are censored, that is, the event of interest has not

occurred by the end of the study. In my case, the event of interest is remediation. Unfortunately, it is quite common for resources to go unfixed and therefore to be censored. The value and sign of each coefficient quantifies the impact of a variable on the instantaneous *hazard* ⊛ of incidents surviving beyond a point in time, commonly referred to as the hazard function $h(t)$. This is relative to a baseline hazard derived from observations. This function is specified as $h(t, X) = h_0(t) \times e^{X^T\beta}$, where h_0 is the baseline hazard, $X_{p\times 1}$ a vector of p explanatory variables and $\beta_{p\times 1}$ regression coefficients expressing the direction and magnitude of the effects of each variable to be estimated from the data by maximizing partial-likelihood.

⊛ *The term "hazard" originates from medical studies and refers to hazard of patients dying. Unlike medical studies, higher hazard is preferable in this context.*

Note that my constructed statistical models are non-parametric, i.e., they make no assumption about the particular distribution of the dependent variable, i.e. remediation time. They do, however, assume that the effects of the independent variables remain proportional over the timeline of events. For a detailed discussion of CoxPH models I refer the reader to [148].

Given the dependent variable of remediation time and independent (proxy) variables, I construct and report several CoxPH models over DBL data first. Here, data are reported at the level of individual incidents. Also note that for DBL a more comprehensive set of independent variables are available. To observe the impact of different agents (defenders and attackers) on remediation time, I incrementally add indicators to demonstrate how the resulting causal inferences change when other agents are, or are not, controlled for.

Table 4.3 presents the first two models ⊛. In model m_1, I do *not* control for differences in attacker behavior. In m_2, I control for attacker behavior by introducing dummy variables derived from the `ML` and `AT` proxy indicators. Attackers either use malicious or hacked resources (`ML`) and may use them to host C&C, malware, or phishing domains (`AT`) jointly resulting in 6 strata of attacker behavior. By including 5 dummy variables, I express and control for all strata of attacker behavior.

⊛ *Note that we have repeated observations of incidents for hosting providers, which we cluster on the basis of the name of the provider to which they correspond using a "`cluster`" term in our model specification. This is a necessary step to account for potential correlation among observations and used to estimate robust standard errors (reported in brackets below estimated coefficients) through the procedure described in [148] (Ch. 9).*

The first set of findings from these statistical models concerns the relative explanatory power of the different classes of indicators. Conventional wisdom about web security holds that less popular websites have worse security because they lack the resources of more popular websites to invest adequately in security. Yet a simple model (not reported in Table 4.3) that only includes webmaster characteristics (in particular website popularity captured by proxy of `AR`) explains almost none of the variation in remediation times (pseudo-R^2 value of 0.001). However, relative to the webmaster-only model, we observe a 33%-pt. increase in explained variance for model m_1 that incorporates provider-level behavior. This increased (Cox & Snell) pseudo-R^2 suggests that providers and their security efforts impact remediation, and are much more influential than domain popularity. Model m_2, which additionally

controls for differences in attacker behavior, increases the proportion of explained variance to 0.59. This 26%-pt. increase in pseudo-R^2 from m_1 to m_2 shows that attacker behavior substantially affects remediation times. It is striking that the models, comprised of imperfect and incomplete indicators, nonetheless explains a majority of the variance in the remediation times.

The next set of findings concerns the influence of particular indicators on remediation times. To draw conclusions, I must first briefly explain how the regression coefficients of these models may be interpreted. Consider the coefficient for the `RNG` indicator from model m_1. This indicator captures the size of a hosting provider (see Table 4.2). `RNG` has a significant, positive effect of 0.632 on the hazard rate. This signals an increase in hazard if all other indicators were held constant, and the number of IP addresses assigned to a provider were to increase by one over its mean value. Increased hazard means that incidents are remediated faster. But how much faster?

We have to exponentiate the coefficient to interpret it accurately ($e^{0.632} = 1.88$). Since `RNG` is Log_{10} transformed, we can say that for every 10-fold increase in the number of IP addresses assigned to a provider, remediation speed increases by 88% compared to the baseline. Negative coefficients on the other hand, indicate slower remediation. It is important to note however, that only because m_1 and m_2 share the same baseline hazard their coefficients may be directly compared.

All coefficients are significant in model m_1. In addition to the

Table 4.3: CoxPH Models (DBL)

	Dependent variable:	
	Remediation Time	
Model	(m_1)	(m_2)
RNG	0.632**	0.399**
	(0.123)	(0.099)
DOMs	0.843**	0.344**
	(0.148)	(0.078)
SDOMs	−0.354*	−0.218**
	(0.139)	(0.073)
IPs	−0.678**	−0.314*
	(0.233)	(0.146)
SIPs	0.373*	0.244*
	(0.179)	(0.109)
CAR	−0.833**	−0.296**
	(0.076)	(0.063)
MAR	−0.201**	−0.145*
	(0.066)	(0.072)
SPM	−0.350**	−0.083
	(0.111)	(0.065)
AR	0.104**	−0.037
	(0.024)	(0.028)
Hacked.C2		1.543**
(dummy)		(0.112)
Hacked.Malw		3.526**
(dummy)		(0.152)
Hacked.Phish		4.435**
(dummy)		(0.258)
Malicious.Malw		2.306**
(dummy)		(0.186)
Malicious.Phish		2.246**
(dummy)		(0.168)
Observations	154,052	154,052
pseudo-R^2	0.33	0.59
Log Likelihood	−1,244,755	−1,207,761
Note:	*p<0.05; **p<0.01	

number of IPs assigned to a provider, the number of domains hosted and the number of IPs operating at least 10 domains are linked to faster remediation times. By contrast, the number of shared domains, the number of IPs hosting websites, the number of Alexa ranked websites, and their median Alexa ranking are associated with slower remediations.

Comparing the coefficients of m_1 to m_2, where more differences have been controlled for, we observe that the magnitude and significance of indicators shared across both models are lowered. However, it is worth noting that with the exception of AR (which I have already shown to have weak explanatory power), the significance and direction (positive or negative) of the coefficients is consistent across the models.

By incorporating attacker related indicators in m_2, we can disentangle their effects from the provider-level measures. The coefficients are computed relative to maliciously registered C2! domains. All other attack categories are cleaned up faster. Hacked phishing websites are fastest, followed by hacked malware. Maliciously registered websites, be it for phishing or malware, are cleaned more slowly than their hacked counterparts. More important than the individual differences between categories, what is especially noteworthy is that a lot of the total variation in cleanup times is tied to the decisions taken by attackers, as opposed to anything that the defender, be it webmaster or hosting provider, can do. This is consistent with Çetin et al.'s experiments on notifications involving the Asprox botnet, when they observed that remediation times worsened substantially when the botnet operators took steps to evade detection [21].

While prior work has confirmed the important role of webmasters [135], these results demonstrate that providers also play a very significant role in abuse remediation, not only in preventing the incidents, but also in responding to them once they occur. Nonetheless, the attacker also influences remediation times, as just reported. Note that the results in Table 4.3 are robust and still hold after residual analysis and removing observations with the top 5% largest residuals (i.e., outliers). The results are also consistent with findings from [78], which report that patching efforts by shared hosting providers had a large impact on the number of abuse incidents, even for client-side software.

Triangulation

Thus far, I have identified significant explanatory factors for variations in the remediation time of events captured by the DBL dataset. At this point, the findings only hold for this (admittedly large and influential) dataset.

In order to evaluate the robustness of the findings and to check whether they might extend to other forms of abuse data, I now seek to

triangulate my findings by analyzing data from a completely different source.

I therefore compare the estimated causal impacts of the collected indicators across my two data feeds using a single model m_3 that includes data from GSB and DBL. Table 4.4 presents the results. Because GSB only includes malware abuse data, I only incorporate the best matching subset of our DBL data observations, i. e. those labeled as malware abuse (DBL_{malw}). I control for differences in provider and webmaster behavior using the same indicators as before. I am not however, able to control for differences in attacker behavior in terms of abusing hacked or maliciously registered domains in GSB, as this data is not available for the GSB observations and cannot be collected retrospectively at scale.

To triangulate, I combine the GSB and DBL_{malw} data into a joint larger dataset and introduce a dummy variable (`GSB`) that captures which subset of the data each observation belongs to, and use this combined dataset as the input of m_3. This ensures that coefficients for each subset of the data are estimated relative to the same baseline hazard ⊛.

Table 4.4: CoxPH Model - GSB vs DBL (malware)

	Dependent variable:	
	Remediation Time	
Model (m_3)		Interaction ($GSB \times Var$)
RNG	-0.192^{*}	-0.315
	(0.088)	(0.220)
DOMs	-0.564	-0.258
	(0.357)	(0.498)
SDOMs	0.317	0.758
	(0.316)	(0.527)
IPs	-0.195	1.411^{**}
	(0.170)	(0.503)
SIPs	0.296^{**}	-1.568^{**}
	(0.088)	(0.459)
CAR	0.299^{**}	-0.117
	(0.088)	(0.128)
MAR	-0.432^{**}	0.286
	(0.163)	(0.157)
SPM	0.103	-0.385^{*}
	(0.107)	(0.173)
AR	-0.100	0.079
	(0.058)	(0.083)
GSB (dummy)	-4.308^{**}	
	(1.131)	
Observations	146,316	
pseudo-R^2	0.126	
Log Likelihood	$-1{,}048{,}811$	
Note:	*p<0.05; **p<0.01	

I quantify changes in coefficient values across the two subsets by introducing interactions between variables and the `GSB` dummy. These interaction terms (reported in the third column of Table 4.4) are the primary quantity of interest of m_3. They express how much coefficient values for the different indicators previously related to remediation in m_1 and m_2, differ between the two data sources and whether this difference is significant.

⊛ *It is important to note that the effects of indicators that are in fact modeled, are reported against a different baseline and not directly comparable to results reported in Table 4.3.*

We observe that the majority of interaction coefficients do not have a significant value. This signifies that the causal effects of the corresponding indicators do not differ significantly across the two subsets of data and generalize.

A number of interaction terms in m_3 however, indicate inconsistencies when looking at different subsets of the abuse data. That is, each feed suggests different causal effects for some of the indicators capturing hosting provider behavior, namely `IPs`, `SIPs` and `SPM`. This is observable by the significance of the coefficients of their corresponding interaction terms. Their significance should be interpreted as the causal effects of these indicators behaving differently when looking at the GSB or DBL_{malw} subset of the data.

Discussion

It should not come as a surprise that m_3 suggests some inconsistencies. The fact that m_3 does not adequately control for attacker behavior does not influence the fact that a number of factors that have been explicitly modeled across all models behave differently. The model still quantifies the differences of both datasets independently of which factors have been taken into account. The purpose of this triangulation exercise however, is to test to what extent my empirical findings are sensitive to the idiosyncrasies of each dataset and whether they produce generalizable results.

So what generalized findings can we draw? The GSB model still consistently demonstrates that providers play an important role. This is supported by an increase of 11% in the proportion of explained variance (pseudo-R^2) in models that explicitly model webmaster and provider behavior, relative to a webmaster only model.

I hypothesize that the inconsistencies across my datasets are partially driven by their different natures. Data in each feed is collected for different purposes, combating spam and harmful emails in one case (DBL), and preventing Internet users from browsing drive-by download URLs in the other (GSB). These feeds also vary in how they are disseminated, as well as whether and how the affected webmasters and providers are notified.

Beyond the challenges of triangulation, my modeling efforts may be further improved through the collection of a more comprehensive set of proxies during the study period rather than in retrospect. A variety of indicators for the efforts of different types of defenders cannot be accurately collected retrospectively (e.g. how defenders and attackers behaved at the time of each incident). Another limitation is the multicollinearity of some of the proxy indicators which complicate their interpretation. However, I do not expect such improvements to have a significant bearing on the overall results that we find.

The lessons drawn here are clear. Different agents such as webmasters, providers and attackers causally affect remediation times as demonstrated by the results. It is important to explicitly model and disentangle their behavior when drawing causal inferences from remediation time data and to triangulate results to demonstrate generalizability. To understand exactly how and to what extent these agents causally influence remediation times remains to be further explored.

4.6 RELATED WORK

(Studies of Abuse Concentration) Past empirical studies have demonstrated that higher rates of abuse are correlated with a lack of security hygiene and a lack of proactivity to implement security countermeasures. For example, network operators that do not fix security misconfigurations [85], or webmasterss and hosting providers that do not patch vulnerabilities [78], experience more abuse. Given that such correlations exist, numerous studies, including some of my own which I have discussed in the previous chapters, follow this logic in reverse, and compare defenders based on *abuse concentration* [60, 92, 97, 100, 101]. By comparing abuse concentration across provider networks, such studies imply that providers with higher abuse rates show less security effort and may be identified as lax or potential candidates for intervening against.

Studies that evaluate defenders responses in this way, count the number of incidents attributed to each defender. They are not concerned with how fast incidents are remediated but rather their volume. As such, they mostly evaluate proactive defender security efforts in preventing incidents [43, 100] in contrast to the approach of evaluating reactive efforts towards neutralizing incidents that have occurred as fast as possible. A particular drawback of comparing defenders based on incident counts is that one needs to account for the fact that defenders with more infrastructure or customers will experience more incidents [97, 101]. The approach presented in this chapter does not suffer from this particular restriction.

At the same time evaluating and comparing defenders based on incident counts, shares several challenges with comparisons based on remediation times. While the latter involves an additional technical challenge of tracking incidents over time, variations in both incident counts and remediation time are driven by multiple entangled causal factors such as webmaster, hosting provider and attacker behavior [43, 53, 97]. There is also the shared problem of how to deal with noisy and error prone industry data to compare defenders [101]. A handful of prior studies [78, 97] have analytically discussed various types of measurement errors and methods of dealing with multi-causality when dealing with incident counts. To the best of my knowledge however,

methodologies for dealing with these issues regarding remediation times have not been adequately addressed in prior work. This chapter partly fills this knowledge gap.

(Studies of Remediation Time) While most studies that compare defenders draw comparison based abuse concentration, a wide range of studies nevertheless measure remediation times. i mostly analyzed these studies to enumerate different types of measurement errors in remediation times. These studies however, are mostly concerned with the effectiveness of specific countermeasures and interventions, and study the effect they have on how long it takes to remediate incidents [19, 21, 30, 60, 135, 141, 144] or patch vulnerabilities [54, 77, 136, 140, 149] after specific treatments have been introduced. As such, they compare the same defender entity pre and post the introduction of a treatment. Some also examine overall market trends or the responses of a group of subjects to the treatment of interest. By contrast, the study presented in this chapter is concerned with evaluating and comparing defenders against others, and is not concerned with the effect of a specific treatment. Notably several studies are similar to this one and compare defenders against others based on remediation times. These however, have examined the role of other classes of defenders, for example TLD operators [112] or compared hosting provider responses with respect to a specific type of abuse only [3].

Of these prior studies, a few follow a randomized control experimental design [19, 21, 54] to measure the effects of treatments on remediation time. Others follow quasi-experimental designs [77, 135, 136, 141, 149] or are purely observational by nature [30, 60, 144]. While variations in remediation time may be more safely interpreted as causally driven by the introduced treatments in randomized studies, conducting such studies are not always viable options. This study, and several prior studies of remediation times are observational in nature. To this end, this chapter is a first to highlight some of factors that causally drive remediation time in addition to investigate various types of measurement errors in industry abuse data, which need to be accounted for in observational studies. I demonstrate a modeling approach for dealing with this multi-causality challenge, and a triangulation approach for dealing with measurement error, which prior studies have not specifically undertaken and described as limitations.

4.7 CONCLUDING REMARKS

This chapter has investigated how to leverage industry sources of abuse data involving websites to evaluate the reactive security efforts of defenders. Rather than causally attribute the variance in remediation times derived from abuse data to a specific actor, I have modeled the

behavior of multiple actors (attackers, webmasters, hosting providers) and analyzed their impact jointly.

I found that hosting providers significantly influence remediation time, explaining between 12% of the variation for GSB data and 33% of the variance in DBL data. Attacker behavior explained a further 26% of the variance in the DBL data. Hence, I can confidently conclude that providers have a substantial role to play in remediating online abuse. Furthermore, I have demonstrated that my results are partially robust by triangulating the statistical results based on the GSB data against those based on the most comparable subset of DBL. Prior studies examining remediation times have not explicitly triangulated data in this way. Alas, the results also uncover a number of inconsistencies in the direction and significance of certain provider-level indicators of their security behavior. While disappointing, it is worth noting that these datasets are collected and disseminated for different purposes, such as spam filtering or protecting web browser users. Therefore, it should not come as a surprise that some findings do not generalize across datasets. At least now we have 'known unknowns', rather than an overconfident interpretation of a model based on a single dataset. One way to draw more generalizable inferences is to control for differences among the data generation processes. Unfortunately, this is not possible for industry-sourced datasets. A more realistic alternative is to triangulate with other datasets.

I have also demonstrated that abuse data sources present a variety of measurement challenges for researchers investigating remediation times. I uncovered several types of measurement errors in the data (e.g., abrupt drops in survival probability that are consistent with what our industry sources, Google and Spamhaus, document about their data collection and maintenance procedures). I carefully unravel these, illustrating how such errors complicate the construction of metrics to evaluate and compare the remediation efforts of defenders. Different metric specifications constructed over the same dataset can easily lead to conflicting inferences about how the efforts of defenders compare to each other. Of importance is to analyze and construct metrics that are less sensitive to such errors. Metrics that more transparently retain the variations captured by the data should be preferred. Many observed measurement errors reflect conscious choices by the industry source to optimize operational efficiency over the precision of secondary measurements such as remediation time.

Furthermore, the exact data generation process is often purposefully underspecified to prevent gaming by attackers. Consequently, over the long term, researchers should consider additional options beyond constructing more robust metrics alone.

One option is for researchers to take control over the data generation process when collaboration with the source. For example, industry

sources could aid researchers by monitoring abuse remediation frequently and uniformly across networks thereby reducing measurement artifacts and removing systematic biases such as monitoring some networks more than others. A second option is for researchers to reverse engineer the data generation black box. The biases that are identified this way would allow for the use of custom-tailored statistical methods to reduce their impact on the analysis. However, such efforts are likely to be expensive, and measurement error could only be reduced, not eliminated. The final option (and the one adopted by our study) is to triangulate with other data sources, to better identify which findings generalize and which apply only to particular circumstances.

In sum, I hope that these lessons and recommendations help future security research based on remediation times of abuse events and vulnerabilities. A key challenge for security researchers is to improve our understanding of the causal factors that drive security or insecurity. Any study that observes a change in attack, incident, or vulnerability data and then attributes this change to a certain countermeasure or defender action, is making a causal claim. The analysis presented here has aimed to contribute to a better foundation for such claims.

5 THE CASE OF BULLET-PROOF HOSTING

This chapter focuses on examining so-called Bullet-Proof Hosting (BPH) providers, which have been a difficult area of the hosting market to tackle. BPH providers knowingly allow miscreants to host harmful content and abuse their services. They even assist in the persistence of the harmful content hosted on their platform thereby enabling a large range of cybercrime and providing a stable environment for cybercriminals to conduct illicit activities.

BPH providers are interesting cases to examine as they demonstrate the limitations of security metrics. They represents corner cases for which security metrics begin to fail, and exemplify how metrics, including ones that I have discussed in previous chapters, may be gamed and distorted.

Understanding how such malicious providers operate is an important issue due to the pivotal role that BPH plays in enabling cybercrime and may lead to better techniques for detecting and disrupting their operations. Therefore within this chapter I undertake a case-study of a recently taken-down BPH provider called `MaxiDed`. The study on which it is based [103] provides a wide range of empirical insights into the inner workings of BPH providers and demonstrates with ground-truth data how isolating criminal areas of the hosting market via employing security metrics is a challenging task.

My analysis of `MaxiDed` is based on data extracted from its backend databases after it was legally taken down by law enforcement and its servers seized. I connect the extracted data to various external data sources and characterize `MaxiDed`'s business model, supply chain, customers and finances. I reason about what the "inside" view reveals about potential chokepoints for disrupting BPH providers as well as demonstrate how little of the harmful content hosted on the provider's platform is captured in abuse data. As such security metrics that rely on abuse data are unable to correctly capture and represent the concentration of abuse around BPH providers.

5.1 INTRODUCTION

Bullet-Proof Hosting (BPH) is a part of the hosting market where its operators knowingly enable miscreants to serve abusive content and actively assist in its persistence. BPH enables criminals to host some of their most valuable resources, such as botnet Command-and-Control (C&C) assets, exploit-kits, phishing websites, drop sites, or even host child sexual abuse material [24, 150, 52, 143, 37]. The name refers to

the fact that BPH provides "body armor" to protect miscreants against interventions and takedown efforts by defenders and law enforcement.

Much of the prior work in this area has focused on how to identify such malicious providers. Initially, BPH providers served miscreants directly from their own networks, even though this associated them with high levels of abuse. Famous examples of such providers include `McColo Corp.` [151], the `Russian Business Network (RBN)` [152], `Troyak` [52] and `Freedom Hosting` [153]. This operational model enabled AS-reputation based defenses, such as Fire [36], BGP Ranking [105] and ASwatch [93]. These defenses would identify networks with unusually high concentrations of abuse as evidence for the complicity of the network owner, and thus of BPH.

AS-reputation defenses became largely ineffective when a more "agile" form of BPH emerged. In this new form, providers would rent and resell infrastructure from various legitimate upstream providers, rather than operate their own "monolithic" network. Concentrations of abuse were diluted beyond detection thresholds by mixing it with the legitimate traffic from the ASes of the upstream providers.

In response, researchers developed a new detection approach, which searched for concentrations of abuse in sub-allocated IP blocks of legitimate providers [37, 143]. This approach assumes that honest upstream providers update their WHOIS records when they delegate a network block to resellers. It also assumes that the BPH operator functions as a reseller of the upstream providers.

A key limitation of this prior work is that it is based on external measurements. This means that we have little inside knowledge of how BPH operations are actually run and whether assumptions behind the most recent detection approaches are valid. A second, and related, limitation is the lack of ground-truth data on the actions of the provider. There are minor exceptions, but even those studies contain highly sparse and partial ground-truth data [37, 150].

This chapter presents the first empirical study of BPH based on comprehensive internal ground-truth data. The data pertains to a provider called `MaxiDed`, a significant player in the BPH market. It unearths a further, and previously unknown, evolution in the provisioning of BPH, namely a shift towards platforms. Rather than `MaxiDed` renting and reselling upstream resources on its own, it offered a platform where external merchants could offer, for a fee, servers of upstream providers to `MaxiDed` customers, while explicitly indicating what kinds of abuse were allowed. By operating as a platform, `MaxiDed` externalizes to the merchants the cost and risk of acquiring and abusing infrastructure from legitimate upstream providers. The merchants, in turn, externalize the risk of customer acquisition, contact and payment handling to the marketplace. This new BPH model is capable of evading the state-of-the-art detection methods. Our analysis shows that in most cases, there

are no sub-allocations visible in WHOIS that can be used to detect abuse concentrations, rendering the most recent detection method [37] much less effective.

Before we can develop better detection and mitigation strategies, we need an in-depth empirical understanding of how this type of provider operates and what potential chokepoints it has. To this end, I analyze a unique dataset captured during the takedown of `MaxiDed` by Dutch and Thai law enforcement agencies in May 2018 [154]. The confiscated data includes over seven years of records (Jan 2011 – May 2018) on server packages on offer, transactions with customers, provisioned servers, customer tickets, pricing, and payment instruments. In addition to the confiscated systems, two men were arrested: allegedly the owner and admin of `MaxiDed`.

The central question of this chapter is: *how can we characterize the anatomy and economics of an agile BPH provider and what are its potential chokepoints for disruption*? I first describe how the supply chain is set up. Then, I characterize and quantify the supply, demand, revenue, payment instruments and profits of the BPH services offered by `MaxiDed`. All of this will be analyzed longitudinally over seven years. I also explore what `MaxiDed`'s customers used servers for.

The main contributions may be summarized as follows:

- I provide the first detailed empirical study of the anatomy and economics of an agile BPH provider based on ground-truth data.
- I map the supply of BPH services and find a highly diversified ecosystem of 394 abused upstream providers.
- Contrary to conventional wisdom, I find that the provider's BP services are not expensive and priced at a 40-54 % markup to technically similar non-BP offers.
- I quantify demand for BPH services and find it resulting in a revenue of 3.4M USD over 7 years. I conclude the market to be constrained by demand, not by supply, i.e. demand for this type of agile BPH seems limited.
- I estimate profits to amount to significantly less than 280*K* USD over 7 years. This belies the conventional wisdom of BPH being a very lucrative business.
- I find disruptable pressure points to be limited. Payment instruments were sensitive to disruption, but a recent shift to cryptocurrencies limits this option. I identified 2 merchants and a set of 15 abused upstream hosting providers as pressure points though their identifcation would have been difficult based on external measurements. The only remaining viable options are raising operational costs and taking down the provider's platform.

I should note that the "bullet-proof" metaphor seems less suited for this new model of BPH provider that we study. Commonly, BPH is understood to include two aspects: (i) intentionally enabling abuse, and (ii) providing resilience against takedowns. The BP metaphor directs attention to the resilience. This new business model, however, primarily focuses on the agile enabling of abuse at low cost. `MaxiDed` and its external merchants provide servers for abuse at close to the market price for legitimate servers. Customers then prepay the rent for these servers. This means that the risk of takedown, in terms of a prepaid server being prematurely shut down by the upstream provider, is borne by the customer. Most customers manage this risk by opting for short lease times and treating servers as disposable and cheaply replaceable resources. They take care of the resilience of their services themselves, using these disposable resources. Some forms of resilience – e.g., reinstalling an OS and moving files to a new server – are provided by the BPH provider as a premium service for an additional fee. The 'bullet-proof' metaphor is less suitable for this business model. A more fitting alternative may be "agile abuse enabler". That being said, in this chapter I retain the existing term. The market of intentionally provisioning hosting services for criminals is still widely referred to as BPH and I want to maintain the connection with prior work.

The remainder of this chapter is structured as follows. First, I provide a high-level overview of `MaxiDed` 's business (Section 5.2). I then discuss the ethical issues related to our study (Section 5.3). Next, I describe our datasets (Section 5.4) and the integrity checks I performed to ensure the validity of my analysis (Section 5.5). I then outline `MaxiDed`'s anatomy and business model (Section 5.6). Next, I turn to the substantive findings and analyze the supply and demand around `MaxiDed`'s platform, with a specific focus on identifying choke points (Section 5.7). I also analyze `MaxiDed`'s customer population (Section 5.8). I then take a look at longitudinal patterns in terms of use and abuse of BP servers by customers (Section 5.9). The final part of the analysis is on `MaxiDed`'s revenue, costs and profits (Section 5.10). I conclude by locating this study within the related work (Section 5.11) and by discussing its implications for the problem of BPH (Section 5.13). Additional material are provided in Section 5.14.

5.2 BACKGROUND

`MaxiDed` Ltd. was a hosting company legally registered in the Commonwealth of Dominica, an island state in the West Indies that is also known for its offshore banking and payments processing companies. `MaxiDed`'s operators publicly advertised the fact that customers were allowed to conduct certain abusive activities upon purchasing its hosting solutions. While WHOIS information of the `MaxiDed` domain shows

Table 5.1: MaxiDed in comparison with previously studied BPH by Alrwais et al.[37] that appear to be still operational

	Advertised BPH Services			
BPH	Dedicated Servers	VPS	Shared Hosting	Total
66host	0	0	3	3
outlawservers	1	6	4	11
abusehosting	47	5	3	55
bpw	5	4	0	9
bulletproof-web	7	9	0	16
MaxiDed	1,855	1,066	0	2,921

that it has existed since 2008, web archive data suggest that initially it was just a small hosting provider with no mention of allowing illicit activities. It underwent a major transformation in 2011 towards becoming an agile BPH service. MaxiDed does not have its own Autonomous System (AS), nor does it have any IP address ranges assigned to it by Regional Internet Registries (RIRs), according to our analysis of WHOIS data at the time of its disruption. This implies that IP addresses are provisioned to customer servers by upstream providers, rather than by MaxiDed. This underlines MaxiDed's agile nature, i.e., its reliance on reselling upstream infrastructure. Table 5.1 compares MaxiDed with several previously studied agile BPH providers in terms of the quantity and types of services they offered. It highlights that its scale of operations is around two orders of magnitude larger. It is reasonable to view the provider as a major player in this market which others have similarly pointed to [155].

5.3 ETHICS

Our data is similar in nature to that used in prior studies of criminal backends [26, 29, 156]. It originates from legal law enforcement procedures to seize infrastructure. Using such data raises ethical issues. For this study I operated in compliance with and under the approval of our institution's IRB. I discuss further issues using the principles identified in the *Menlo Report* [157].

(**Respect for persons.**) The data contains personally identifiable information (PII) on customers, merchants and employees. Access has been controlled and limited to authorized personnel within the investigative team, and later granted to several of the co-authors. Since 'participation' in this study is not voluntary and cannot be based on informed consent, we took great care not to analyze PII on customers, because they form the most vulnerable party involved and not all of them may have

used servers for illicit purposes. I only compiled aggregate statistics. For merchants, I have masked identities using pseudonyms to prevent identifiability. I did not analyze the data in terms of `MaxiDed` employee names.

(**Beneficence.**) I believe that our analysis does not create further harm. I did not purchase services from the provider and thus did not contribute to any criminal revenue. The authors and police investigators believe the benefits of a better understanding of BPH operations, most notably in terms of better countermeasures, outweigh the potential cost of making this kind of knowledge more widely known, as the model of agile BPH itself is already well-documented in prior work.

(**Justice.**) The benefits of the work are distributed to the wider public, in terms of helping to reduce crime. It especially helps to protect persons who are more vulnerable to being victimized. I see no impact to persons from being included in the study itself.

(**Respect for law and public interest.**) This study has been conducted with the approval of, and in collaboration with, the investigative team and public prosecutors. It is important to note, that while captured information may point to certain illegal conduct, establishing legal proof of criminal conduct is *not* the purpose of this study.

5.4 DATA

Table 5.2: High-level statistics of `MaxiDed` backend data

Data on	Description	Total Nr.
Suppliers	60 directly listed upstream hosters and 14 listed merchants supplying server packages	74
Server Packages	Customizable server packages on offer during 2011-2018	56,113
Payment Instruments	Supported payment instruments/methods	23
Orders	Customer placed orders for various server packages and other administrative services	66,886
Users	Number of registered users	308,396
Transactions	Financial transactions including 30938 received payments and 33124 payments made to other entities	64,602
Tickets	CRM system tickets capturing communications between various entities	26,562

From the servers seized during the takedown, the Dutch investigative team has been able to resurrect `MaxiDed`'s administrative backend (CRM and database). They have granted us access to the data and corresponding source code. I analyzed the source code to ensure correct interpretation of the stored data. I observed how various resurrected administrative pages queried specific records to display information.

The revived single-instance `Postgres` database contains longitudinal information on several key aspects of `MaxiDed`'s operations. On the supply side, it includes data on what server packages were on offer, which merchants were offering these packages, and the internal and externally-advertised prices of each package. On the demand

side, there is customer contact information, order placements, rented servers, server assigned IP addresses, financial transactions, and type of payment instruments used and available over time.

Communications between MaxiDed operators, customers, merchants, and upstream providers were captured as CRM system tickets. Ticket contents and email communications also include instances of abuse complaint emails that MaxiDed administrators received and forwarded to their customers. I should note that the operators also operated a live-chat channel for customers on the site. They were also known to use ICQ, Jabber and Skype contact channels at some point in time. These communications were not stored on the seized servers, if they were stored at all. Communications data, often the most sensitive, have not been analyzed in favor of the ethical principles that were followed.

Overall, the retrieved data represents information over the course of MaxiDed's life span from Jan.- 2011 to May-2018, when its operation was disrupted. High level statistics and descriptions of the ground-truth data is presented in Table 5.2.

To enrich the ground-truth data, I deployed several additional data sources. Domain-based resources operating from the customer IPs, were identified using historical passive DNS data collected via Farsight Security's (DNSDB [111]). To identify upstream providers of servers and IPs, we used historical acsWHOIS IP allocation data from Maxmind [107]. A set of domain and IP-based blacklists have been used to gain further insights into abuse emanating from customer servers.

5.5 DATA INTEGRITY

Since we did not gather the information ourselves, there is a need to evaluate its accuracy and authenticity: how do we know that MaxiDed admins did not manipulate data, for reasons of operational security or otherwise?

The data resulted from the legal seizure of servers, in close coordination with apprehension of two individuals who had administrative control over these systems. This ensured that the data was not manipulated during or after the seizure. To ensure that data was not manipulated in the course of MaxiDed's operation, I have examined data integrity in several ways. I first discuss the correspondence of the seized data with external (third-party) data. Next, I analyze the internal consistency of the seized data itself.

The strongest indicator of integrity is that the seized server data was consistent with the data that was collected via legal intercept prior to the takedown. A wiretap had been running for over two years on the backend CRM server.

I also compared the data to snapshots of MaxiDed's webshop archives on Internet Archive between 2015-2018. I extracted all server package

IDs that were on offer. All these IDs were present in the seized back-end data as well.

For a sample of over 50 server packages on sale in April 2018, I compared the internally recorded price with the prices of the entities listed as the upstream providers. These included packages from a Dutch and a German upstream hosting provider. For each package, I visited the supplier's website, customized a server package to match, and found its price to be correctly reflected by the internal price.

For the payment data, I was able to compare the WebMoney transactions logged in the database with data that was subpoenaed by Dutch law enforcement from WebMoney on transactions during a period of 10 days involving one particular WebMoney wallet address. Of 31 internally recorded transactions during this period via WebMoney, 17 were matched with the external data.

Together, these external checks provide confidence that the internal data has not been manipulated. Multiple *internal* data consistency checks were also carried out. I cross referenced customer order placements against server package data, to determine if all order placements consistently point to an existing package. Of the 14,702 customer orders for servers, I found 431 referencing package IDs that were not listed, indicating a 2.9% proportion of inconsistent order placement records. These references point to a set of 306 unique server packages (a 0.5% proportion of all server packages).

I also cross referenced MaxiDed operators' payments to their merchants, against server package data. These indirectly referenced specific server packages, thereby indicating what each payment is for. Of the 33,124 outgoing payments, I found 345 referencing packages that were not listed among the set of offered server packages (a 1.0% proportion of inconsistent payment records). Cross referencing the same payment data against customer orders, I found 474 outgoing payments referencing servers that were not listed among the orders of customers (a 1.5% of inconsistent payment records).

The timestamps of order placement and transactions were also analyzed, to check for suspicious gaps in the timeline. The longest gap was observed to be 76 days from 2011-03-31 to 2011-06-15. All remaining gaps (37) were at most 2 days long. Approximately an average number of 26 order placements per day were observed. For payment events, the longest timeline gap was observed to be 135 days pertaining to the data from the period between 2011-01-29 and 2011-06-13. The remaining gaps (5) were no longer than 1 day. An average number of 24 transactions per day were observed in the payment data.

The minor inconsistencies and timeline gaps for the most part relate to records from 2011 and 2012, a period corresponding to the initial set up and early growth phase of MaxiDed. A certain amount of inconsistency in database records is to be expected, but more so during

the initial set up and growth phase of any organization. All in all, the internal and external consistency of the data merits confidence in its validity for the purposes of characterizing the overall anatomy and economics of MaxiDed's BPH operation.

5.6 ANATOMY OF maxided'S BUSINESS

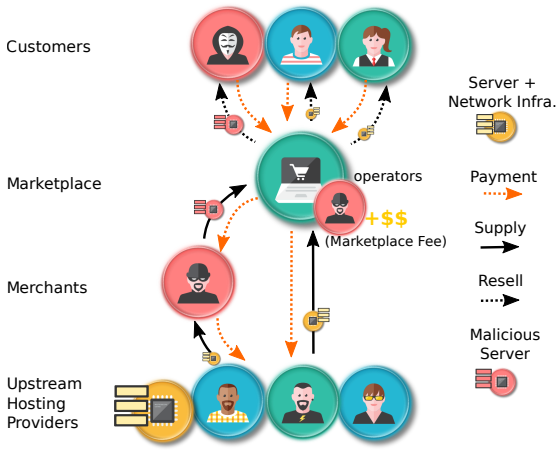


Figure 5.1: MaxiDed in a glance.

Figure 5.1 provides a high-level overview of MaxiDed's anatomy and business model. In following sections I will take a close look at each of its components.

Hosting Business Components

(**Marketplace**) MaxiDed was a marketplace which connected merchants offering server packages that allowed abuse, with customers looking for an abuse-tolerant provider. It captured a fixed 20% fee from each sale between a merchant and a customer. Customers did not see the merchants' identities or even that an offer came from a separate entity. All they knew was that they contracted with MaxiDed. The merchants advertised server packages from legitimate upstream providers and put these on the MaxiDed market with a markup. Server packages specified default server configurations that were further customizable by customers. In addition to the technical specification, each package indicated what type of abuse, if any, was allowed. The majority of the packages explicitly allowed certain forms of abuse. MaxiDed itself also put server packages from certain upstream providers for sale in the webshop, de facto operating as merchant on its own platform.

For its own packages, profits varied between 0 to 40% of the cost of packages at the upstream providers. What's more, MaxiDed also operated as a customer on its own platform, acquiring offers from merchants for its side business, a highly permissive and lucrative file sharing service called DepFile. This file sharing service was a major hub for distributing child sexual abuse material.

The platform approach means MaxiDed can externalize the cost and risks of acquiring and supplying upstream server infrastructure to third-party merchants. As such it is decoupled from the upstreams. The advantage for merchants, on the other hand, was that they could externalize the responsibility and risks of acquiring customers and processing their payments. Beside the fee that MaxiDed charged on top of the merchant's price, it also charged customers for performing additional administrative tasks, like re-installing servers after a takedown by the upstream provider. From these fees, it needed to recoup the cost of its staff and backend systems.

The main components of the marketplace were a frontend webshop, a backend Customer Relationship Management (CRM) system, accounts for merchants who could offer server packages on in the webshop, and payment handling of customers paying to MaxiDed and, in turn, MaxiDed paying the merchants when their offers resulted in a sale. The CRM, a series of webpages implemented in PHP, was used by both MaxiDed and merchants to create the server packages displayed on the webshop. It was also used to facilitate communications between customers and merchants through customer tickets. Merchants were responsible for handling customer tickets of their own server packages. Communications also took place through multiple MaxiDed support email addresses which were automatically imported into the backend database and live-chat functionality which was not retrievable from our data.

Different payment options have been supported over time by MaxiDed; 23 in total. Some from third-party payment providers like Paypal and WebMoney to cryptocurrencies such as Bitcoin and Zcash.

(**Merchants**) Third-party merchants supplied server packages that were re-branded and sold, with a mark-up, under MaxiDed's name. Many offered packages were directly scraped by the merchants from retail auction sites run by certain upstream providers. As far as I could tell, most merchants had no established reseller relationship with the upstream provider and no delegation was visible in IP WHOIS (I explore this more systematically in Section 5.7.3.). This invalidates a key assumption in prior work, i.e., that agile BPH providers operate on the basis of established reseller relationships that are visible in sub-allocations. In some cases, merchants did establish reseller relationships with an upstream provider. This allowed them to hook into an API and automate the importing and advertising process of upstream pack-

ages, rather than having to manually scrape other hosting provider's websites, in addition to receive certain discounts.

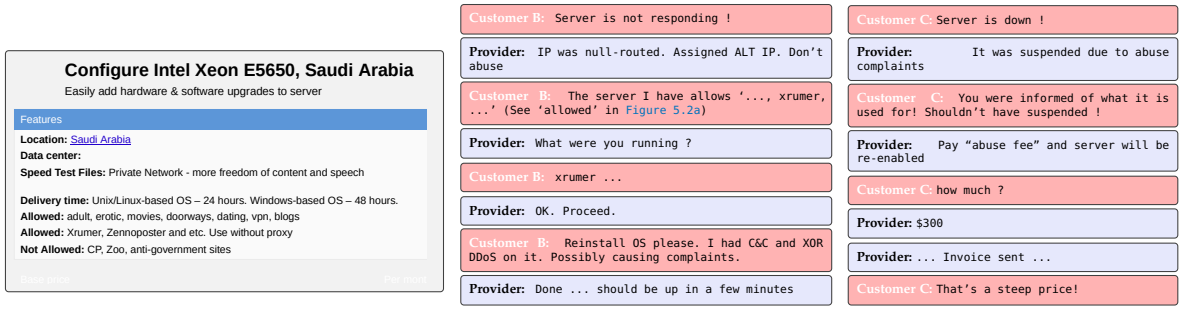
(**Upstream Providers**) These are legitimate hosting companies that offer server packages, via retail channels, auctions or reseller programs, which are put into the MaxiDed marketplace by the merchants. Once sold, the merchant acquires the package from the upstream provider. In Section 5.7.3, I use WHOIS IP allocation information to infer from which upstream providers the merchants bought their packages.

(**Customers**) Customers were elicited for their preferences and guided towards server packages upon visiting MaxiDed's webshop. This occurred via standard search filters or via live chat with administrators. Customers were able to request more powerful hardware, additional IP addresses, pre-installation of a specific OS, and decide on the physical location of the servers. Figure 5.15 (see Section 5.14 Appendix-A) provides an excerpt of a live chat conducted by one of the authors with MaxiDed operators prior to its takedown demonstrating this process.

Customers would first deposit funds into a USD denominated "wallet" and then use these wallet funds to pay for the invoices that MaxiDed issued to them. In other words, purchases were prepaid. This structure allows merchants to place orders only after receiving payments and to shift the risks of premature contract termination to customers as they have received payments in full. Customers were not reimbursed for lost server-day usage due to premature service suspension at the upstream.

Side Business

MaxiDed's administrators also operated a file sharing platform, known as DepFile [155, 158], run on servers which they rented through the MaxiDed marketplace. Some of these servers were also seized during the law enforcement action. Data shows that DepFile infrastructure was acquired using a single MaxiDed customer account which never paid its invoices. Over time, the account accrued approximately 400,000 USD in debt. DepFile allowed its customers to host and access content, some of which included child sexual abuse material, on a monthly subscription basis. Our separate analysis of internal DepFile data, suggest that it resembled a so called "affiliate program" [13, 29, 159] with affiliates bringing in new subscribers. The profits from subsequent sign-ups were shared between DepFile (a.k.a. MaxiDed) and the affiliates. As an aside: these profits were much higher than those of MaxiDed. One could argue that the MaxiDed was more valuable to its owners as a way to acquire cheap and risk-free server infrastructure than as its own profit model.



(a) (b) (c)

Figure 5.2: Examples of MaxiDed's bullet-proof behavior. (a) screenshot of server publicly advertised to customers. (b) and (c) are excerpts of a conversation between customer and administrator (edited for readability).

Examples of Bullet-Proof Behavior

Figure 5.2a shows a screenshot of one of MaxiDed's publicly advertised server packages along with descriptions of its location, network/IP-address information, price, in addition to explicit descriptions of abusive activities that were (dis-)allowed upon purchasing. Figure 5.2b illustrates a conversation (lightly edited for spelling) that took place between an admin and a customer in the context of a CRM ticket. XRumer is a tool aimed at boosting search engine rankings by auto-registering accounts and posting link spam. It demonstrates that MaxiDed operators were not only explicitly tolerating abuse, but that they were informed about the abusive activities of their customers and actively supported them. This is also the case for DepFile. It knows the file sharing service is supporting illegal content, including child sexual abuse material. The customer interaction also shows the admin ignoring abuse complaints, then assisting the customer by migrating resources to a different network location. Figure 5.2c is another example of a (lightly-edited) conversation excerpt, demonstrating that certain customers were asked to pay an 'abuse fee' to continue accessing their rented server upon receiving abuse complaints.

5.7 SUPPLY AND DEMAND FOR BPH

MaxiDed's operations deviate from certain assumptions underlying recent detection techniques. This warrants a more detailed analysis of its characteristics to understand if this new form of agile BPH exhibits chokepoints that allow for disruption. Most disruption strategies rely either on taking down the provider as a whole or on cutting off the

supply of resources that it needs: servers, connectivity, payment instruments, customers. In MaxiDed's case, the former occurred. These kinds of takedowns however, are rare and hard to scale. This section explores the alternative strategy: squeezing potential chokepoints in the supply chain.

Merchants

In a period of seven years, merchants offered 56,113 different server packages. Around a quarter of all packages (14,931) explicitly allowed certain kinds of abuse. I refer to these as bullet-proof (BP) packages. Note that non-BP packages were also abused, as we learned from customer tickets when servers were suspended. Admins frowned on this practice. Not because of the abuse itself, but because these customers should have purchased a more expensive abuse-allowing package. MaxiDed admins listed offers as well in the role of a merchant on their own platform. We label MaxiDed as *merchant zero* (mc_0) and 14 third-party merchants as $mc_{1...14}$, identified by connecting MaxiDed's user and supplier database tables.

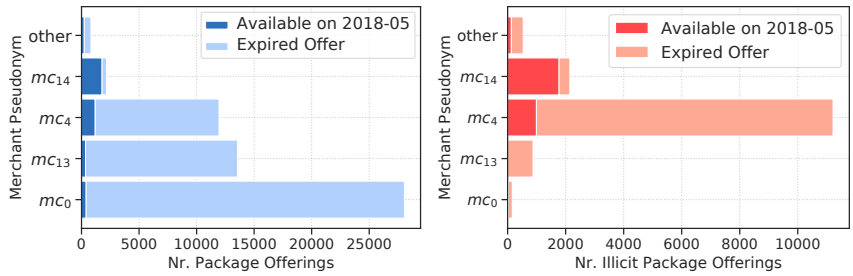


Figure 5.3: Merchant Package Offerings. (**left**) All packages; (**right**) Subset of illicit packages

Figure 5.3 (left) illustrates the total number of server packages offered by the top 4 merchants, which accounted for 98% of all packages. At the moment of takedown (May 2018), there were 3,957 available packages. Of these, 2,921 (74%) explicitly allowed abuse. Packages expired when corresponding upstream provider packages expired or when operators no longer maintained relationships with the upstreams.

Figure 5.3 (right) shows the subset of server packages that allowed abuse, from the same top four merchants. This figure highlights that two merchants, mc_4 and mc_{14} were responsible for 89% of all the BP packages offered on MaxiDed's platform and 94% of the BP packages available at the moment of the takedown. Interestingly, MaxiDed itself (mc_0) supplied only 29 BP packages (1%), relying almost exclusively on its merchants to supply BP infrastructure. This fits with the interpreta-

tion that moving to a platform model allowed MaxiDed to externalize the risk and cost of managing the relationships with upstream providers around abusive practices.

Of the 14,931 BP packages on offer, only 3,066 (20%) were ever sold. There were 9,439 customer orders for these. This indicates that there was an oversupply of BP packages on MaxiDed. Sales followed a similar distribution to supply, with mc_4 and mc_{14} accounting for 70% of all sales. (Of the packages that did not explicitly allow abuse, 2,006 were sold 4,832 times.)

In sum, only around 20% of offers were ever sold, showing that the market for BPH is, unfortunately, not supply-constrained. MaxiDed externalized the supply of BP packages to merchants and two of these were dominant, in terms of supply and sales. Merchants mc_4 and mc_{14} would have been viable candidates for disrupting the supply chain of the marketplace as a whole, had they been identified prior to MaxiDed's takedown. This might be feasible if, as prior work assumed, they are resellers of upstream providers and WHOIS records are updated to show which network blocks are delegated to them. I later discuss evidence that, in most cases, there is no such delegation. The takedown of MaxiDed itself is unlikely to have disrupted these merchants. They may have taken some losses from outstanding due payments from MaxiDed. Except for these losses, merchants could migrate to other marketplaces, resulting in a game of whack-a-mole. This demonstrates the advantages of merchants externalizing part of their risks to the MaxiDed platform.

BP Package Categories

BP packages were differentiated in terms of what types of abuse was allowed. The platform pre-defined 12 categories of abusive activities. Merchants could tick the boxes of whatever categories they were comfortable with for their packages. The activities ranged from the distribution of pornographic content or copyrighted material, to Internet-wide scanning, running counterfeit pharmacies, running automated spamming software such as Xrumer, and doing IP spoofing, typically to conduct amplification Distributed Denial of Service (DDoS) attacks. Table 5.3 lists these activities along with associated category labels $C_{1..12}$.

We suspect merchant choices for certain types of abuse to have been partly driven by what they could handle in terms of their relationship with the upstream provider of a package. Some forms of abuse trigger more backlash than others. Plus, certain upstreams might be less vigilant regarding certain forms of abuse, depending on jurisdiction or other factors.

To analyze the relationships among the allowed forms of abuse, I calculate the correlations between all categories. In other words, if category 'c_X' is allowed, what is the probability that category 'c_Y' is also allowed?

The results are plotted in Figure 5.4. Five groups of server packages can be identified, each with a different type of abuse profile, which roughly corresponds to a certain risk profile. At the top end of the risk profile is "spoofing" ($x = c_{12}$). Where this was allowed, everything else was also allowed with high probability (i.e., all values along the y-axis indicate high probability for $x = c_{12}$). As such a highest risk group label G_5 was assigned to packages that allow "spoofing". One step down are packages that allow "scanning" ($x = c_{11}$): everything else is typically allowed, except "spoofing" ($x = c_{11}, y = c_{12}$), which has a lower probability. This is group G_4. Next, G_3 was assigned to a group composed of 4 categories, $C_{7..10}$ which were allowed in conjunction with a high probability, and disallowed the higher risk $c_{11..12}$ categories with a high probability. The remaining groups were created using a similar logic.

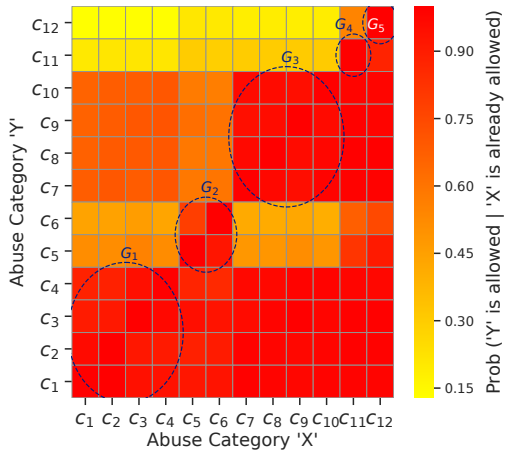


Figure 5.4: Correlation of abuse categories. (See Table 5.3 for c_i labels).

Table 5.3: Statistics on packages allowing each category of illicit activity and associated risk groups

Cat.	Description	All packages	Avail. before takedown	Risk Group	Avail. per-group
C_1	File Sharing	12,344	2,724	G_1	404
C_2	Content Streaming	11,891	2,629		
C_3	WAREZ	11,856	2,615		
C_4	Adult Content	10,732	2,557		
C_5	Double VPN	10,099	1,529	G_2	630
C_6	Seedbox	8,835	1,298		
C_7	Gambling	2,663	1,862	G_3	1,279
C_8	Xrumer	3,120	1,849		
C_9	DMCA ignore	2,978	1,841		
C_{10}	Pharma	2,620	1,821		
C_{11}	Scanning	629	565	G_4	254
C_{12}	Spoofing	396	354	G_5	354

For each risk group, Table 5.3 lists the abuse types and the number of packages that allowed it, over the whole period of `MaxiDed` ('all packages') or at the moment of the takedown ('Avail. before takedown').

Note that packages are counted multiple times, as they often allowed multiple forms of abuse. The last column, 'Avail. per group', counts each package as belonging uniquely to one group, namely the group with the highest risk profile – e.g., if a package allows spoofing, it will be counted in G5, but not in others, even though it likely also allows those types of activities. We can see that `MaxiDed` had a significant amount of supply in each category, with a clear peak in group 3.

A side note: the tickets and live chats clearly showed that other types of abuse were also allowed, such as running botnet Command-and-Control (C&C) servers. The admins did not wish to list these forms of abuse publicly (see Figure 5.15 in Section 5.14 Appendix-A).

Merchant Upstream Providers

To understand how `MaxiDed`'s supply of BP infrastructure was distributed over legitimate upstream providers, I narrowed our analysis to 5 merchants, namely mc_0, mc_4, mc_{10}, mc_{12}, and mc_{14}, who jointly had 94% of the BP package sales.

Merchant mc_{14} sold most of the servers associated with risk groups G_3 or higher, the others sold mostly packages of group G_3 and below. So mc_{14} appears to have specialized in higher risk packages.

I determined each merchant's set of upstream providers by first extracting from the data the IP addresses provisioned once the server was sold. Maxmind's historical IP WHOIS data was then used to lookup organizations to which these IP address belonged. This way, I could see how each merchant's supply chain was composed of multiple upstream providers. The variance was significant. The two dominant merchants (mc_{10} and mc_{14}) abused 134 and 276 upstream providers, respectively. The others connected with 4 to 26 upstreams. Overall, `MaxiDed`'s supply chain comprised of servers at 394 upstream providers.

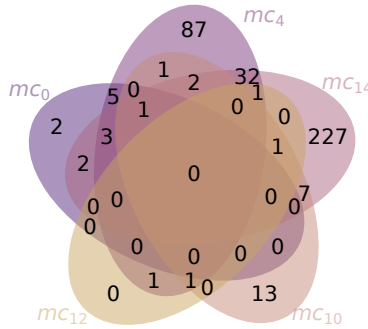


Figure 5.5: Upstream Overlaps

Figure 5.5 show how much, or rather how little, the supply chains of merchants overlapped in terms of upstreams. Figure 5.6 shows a CDF of how each merchant's sold BP servers were distributed across its own set of upstream providers. Across all merchants, 15 upstream hosted 50% of all sold BP servers and 57 account for 80% of all sold servers.

At first glance, the concentration in 15 upstream providers suggests a chokepoint that could be leveraged, but the long tail of available upstreams makes this strategy not very promising. Merchants could shift supply to those hundreds of alternatives. The 15 top ones might

have certain advantages in terms of location, price and quality, but only 5 of them are shared between the two top merchants, so there does not seem to be a unique advantage to these providers.

Recent BPH detection approaches [37] have relied on upstream providers updating WHOIS records when they delegate network blocks to resellers. As stated, `MaxiDed`'s data suggested that merchants often do not enter into reseller agreements with upstream. That would seriously undermine the effectiveness of these detection methods. To test this more systematically, I looked at the set of upstream providers that hosted 80% of the BP servers (57). In this set, I found 22 which are reputable upstream providers and more likely to reflect sub-allocations to their clients in WHOIS. I randomly sampled 10 BP servers for each of these 22 providers and manually inspected their IP WHOIS information. In only 24% of the cases did the WHOIS information reflect sub-allocation to downstream entities. Note that these downstream entities might also be legitimate resellers who sold to the merchants, rather than being the merchants themselves. Also, none of the records pointed to `MaxiDed`. This means that in 76% of the cases, the BP activities could not be associated with a sub-allocation, thus evading the current best detection method. Abuse on these addresses would be counted against the upstream provider, typically diluting the detectable concentration of abuse. Establishing a relationship between the upstream provider, their downstream customers, merchants and, ultimately, `MaxiDed`, would have been impossible with this kind of data.

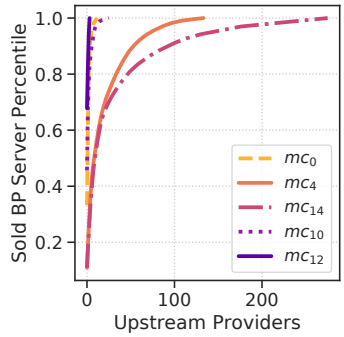


Figure 5.6: BP Server Distribution over Upstream Providers

I next examined the distribution of each merchants' sold BP servers and server life spans across their corresponding upstream providers longitudinally. I visualize some of the results for mc_{14}, who was specialized in selling higher risk BP servers. Figure 5.7 plots the lifespan of mc_{14}'s sold BP servers that allowed "scanning" (left) and "spoofing" (right) for its 10 most misused upstream providers.

Figure 5.7 demonstrates that the merchant's BP customer servers were spatially as well as temporally spread across multiple upstream providers. It also shows that at no point in time, was there a shortage in the supply of servers even for the higher risk server packages. We observe no timeline gap during which servers of a particular group were not provisioned and active. We clearly observe a supply chain that was diversified, yet proportionally concentrated on a limited set of upstream providers. This approach of the merchant seems to be driven by a combination of efficiency in working with a limited set of

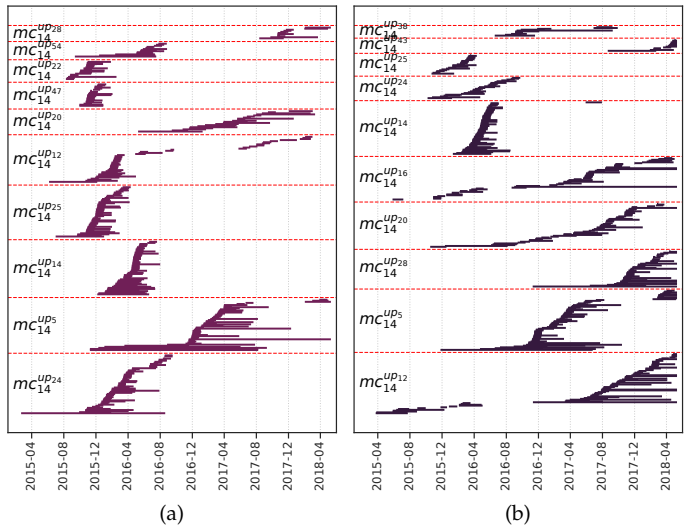


Figure 5.7: 10 most misused upstream providers via which mc_{14} provisioned BP servers of risk group G_4 (allowing "scanning" - left) and G_5 ("spoofing' - right'), plotted against server lifespans at each provider. Each colored line represents the lifespan of one server.

upstreams and the flexibility of migrating from one upstream to the next, once the cost of working with that provider went up, perhaps because of mounting abuse complaints.

Payment Instruments

Next, I analyze the various payment instruments to identify potential chokepoints. From analyzing the source code of the webshop and the transactions in the database, I know that `MaxiDed` accepted payments via 23 different instruments. Three of these were actually never used by customers: `Bitcoin Gold`, `Electroneum` and `Kubera Coin`. Eight payment options were provided for a limited time and then discontinued by `MaxiDed`. At the moment of its takedown, 12 payment options were available. Some of these instruments, e.g., `Paypal`, were later restricted to specific groups of customers. Payments through `Yandex Money` were generally restricted to clients from Russia.

Figure 5.8 reconstructs transaction volumes over time for 20 payment instruments based on timestamps of financial transactions in the data. It plots a logscale of the number of transactions in each month. The Y-axes are the same for all instruments. First, we see that `WebMoney` has been a consistent and reliable payment provider for `MaxiDed`, basically from the start. Other instruments from that period proved more problematic.

For example, Paypal became much more difficult to use in the course of 2015 and was abandoned completely in early 2018.

We can see the operators deploying new ones and also abandoning some of them again. This process seems to suggest responding to potential or manifest disruptions via payment providers. Consistent with this interpretation is the increase in options to pay with cryptocurrencies. We first see a major shift to `bitcoin` at the end of 2013. Then, around the end of 2017, `MaxiDed` added 8 new cryptocurrencies. A preference to move to cryptocurrencies was also observed in backend data, where `MaxiDed`'s operators maintained an explicit preference order for the different payment methods.

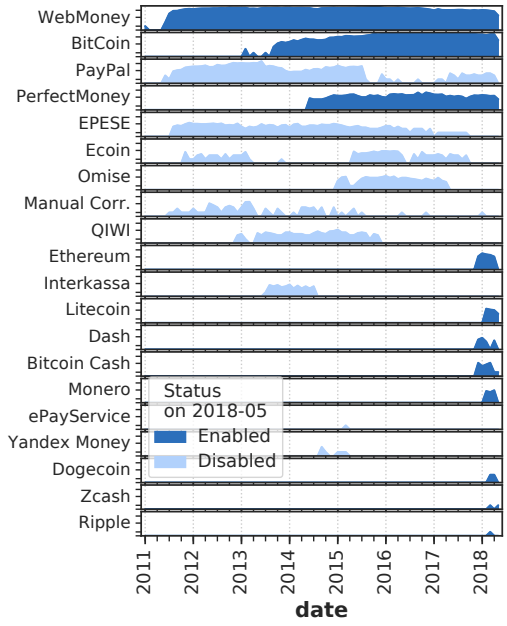


Figure 5.8: Payment instrument monthly transaction volume

Figure 5.9 plots the cumulative generated revenue for the top 5 most popular payment instruments. While `WebMoney` had brought in the most revenue, the total amount of `bitcoin` payments was growing rapidly and poised to overtake the leading position, until the takedown happened.

All in all, `MaxiDed`'s revenue was generated through a small set of payment methods. The bulk of their customers used only one payment method. Disruption of `MaxiDed`'s payment flow via `WebMoney` would have been a viable chokepoint in earlier phases. The self-imposed limits on using `Paypal` probably reflect the fact that those payments were vulnerable to countermeasures by `Paypal`.

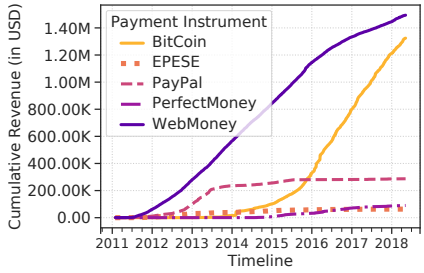


Figure 5.9: `MaxiDed`'s generated revenue per payment instrument

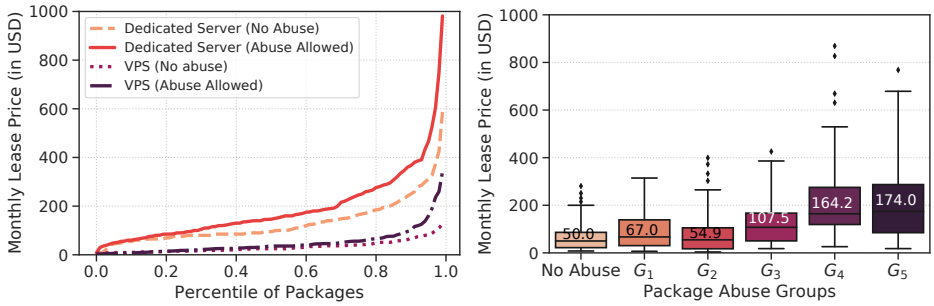
The shift towards cryptocurrency payments demonstrates that `MaxiDed` recognized this dependency, as well as illustrates how it was attempting

to remediate it. It is clear that this shift makes disruption more difficult, though it is hard to gauge how resilient the `bitcoin` payment option actually was. This would require a study of the blockchain and the role of currency exchanges, which is out of scope for this study. That being said, the proliferation of cryptocurrency options might counteract the vulnerabilities associated with each specific instrument.

Package Pricing

BPH businesses are typically understood as charging customers high markup prices for allowing illicit activities and offering protection against takedowns. There is anecdotal evidence (e.g., [150, 37]) that suggests prices are well above those for bonafide services. Our data, however, questions this widely-held understanding.

I first distinguished VPS packages from physical dedicated servers in the data. In each category, I then compared the distribution of the monthly lease price of packages that allowed abuse versus those that did not. The results are plotted in Figure 5.10a. We observe that indeed abuse-enabling servers cost more, but the differences are modest across most of the distribution. For dedicated servers, the median price was 95.00 USD for non-BP packages and 146.00 USD for BP packages. For virtual servers, the median prices were 25.00 USD versus 35.00 USD. These numbers suggest that customers payed a median markup ranging from 40% to 54% for being allowed to abuse. This includes both the fee of `MaxiDed` as well as the margin of the merchant. The rest goes to the upstream provider.



(a) Price per package type

(b) Price per risk group

Figure 5.10: Distribution of `MaxiDed`'s server package pricing (See Table 5.3 for risk group labels).

I also compared package prices based on associated risk groups of their packages. Figure 5.10b illustrates the results with median group prices indicated in the plot. Here, we observe larger prices differences. The median price of the highest risk packages are 3.5 times higher than those for the non-abuse packages.

The limited markup seen in the lower risk packages might reflect the fact that the platform has an oversupply of BP packages. Many packages never got sold. The platform also sets up the merchants to compete with each other. All of this might push prices down, towards the cost of the upstream package. Relatively low markup might also reflect less cost on the side of the merchant and marketplace because of takedown. Low prices may also be the result of `MaxiDed`'s business model which pushes takedown risks to customers by requiring prepayment.

5.8 CUSTOMERS

Law enforcement takedowns of online anonymous markets (a.k.a., dark markets) have targeted the platforms, the supply chains, but also the customers on these platforms, in an attempt to disrupt the demand side. The most ambitious operation was the coordinated Alphabay-Hansa market action, which de-anonymized many merchants and buyers [160]. As of yet, it is unclear if these actions will have any impact on the demand for these services. Nevertheless, I will take a closer look at the population of `MaxiDed` customers to understand how demand has evolved over time and whether it offers starting points for disruption.

`MaxiDed`'s registration data shows that 308,396 unique users signed up to its platform. Figure 5.11 plots the cumulative number of registered, active and paying users over time. I find three outlier events, during which a large number of users appear to have been artificially created, that distort the numbers. Only 6,782 of the user population ever purchased server packages. Of these, 4,498 users were active in the sense that they logged into the platform's CRM at least once after having signed up. On average, the platform saw a daily growth of 3 user sign ups, excluding the three outlier events.

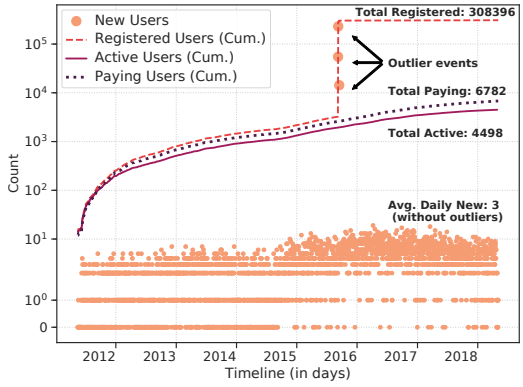
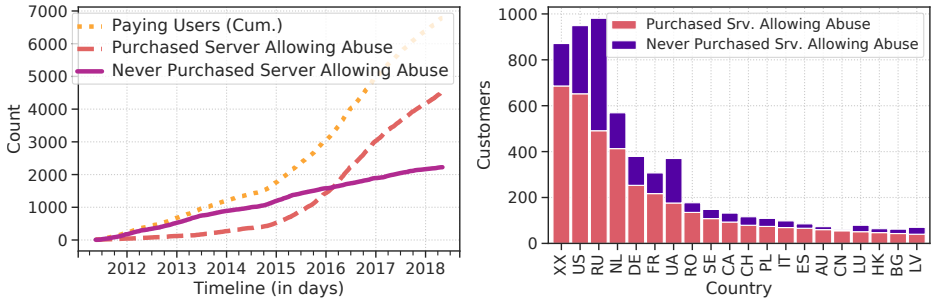


Figure 5.11: `MaxiDed`'s user numbers over time

Cross referencing the user data, customer orders, and server package data, I find that the majority of the customers were interested in and may have engaged in abusive activities. This is observable in Figure 5.12 (left) which plots the cumulative number of customers, separating out those that eventually ended up purchasing BP servers. In the earlier stage of `MaxiDed` 's evolution, they still had a significant number of

customers who never bought BP packages. A few years in, they attract an increasing number of users that do buy BP packages. At the time of its disruption, 66% of all customers ever to register had purchased BP packages. The remaining 34% was a mix of bonafide customers and customers who may have undertaken abusive activities on non-BP packages.



(a) (b)

Figure 5.12: (**left**) Customer types; (**right**) Customer locations (XX = Location not specified)

Customers could specify language preferences in their profile: 5,085 selected English and 1,697 selected Russian. They were also asked to supply location information. Assuming that user-specified locations are correct, a crude assumption, then most users came from 3 countries, namely RU, US and NL (see Figure 5.12 - right), followed by a long tail of other countries.

5.9 USE AND ABUSE

Next, I explore server use and abuse by customers. I examine how customers manage takedown risks transferred to them by `MaxiDed` and look at the measure of last-resort, namely blacklisting BP servers once they are detected.

In Demand Abuse Categories

The data contains timestamps of when servers were provisioned and when they were taken offline. Servers were deactivated when their lease expired or when abuse complaints caused the upstream provider to terminate the lease

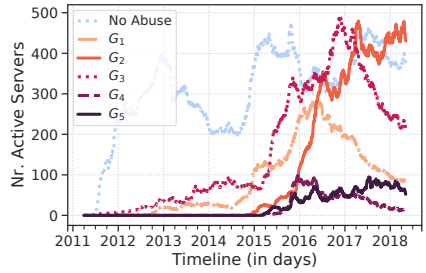


Figure 5.13: `MaxiDed`'s active servers in various risk groups over time

early. Figure 5.13 plots the number of active servers across various risk profiles. It shows what customers mostly sought to purchase.

After a start as a legitimate provider, BP servers become dominant over time. Initially, customers were interested in spamming, operating phishing domains (which triggered DMCA complaints), running counterfeit pharma and gambling sites (risk profile G_3). Then we see a steady growth in demand for G_1: file sharing, streaming, adult content, and WAREZ forums. The rapid growth of `MaxiDed`, starting around the end of 2014, saw a diversification of the abuse and an increase of VPNs and seedboxes for file sharing (G_2), scanning (G_4), and spoofing (G_5). These shifts reflect a wider trend towards commoditization of cybercrime services, such as the provisioning of DDoS-as-a-Service [24]. At its peak, `MaxiDed` administered 1,620 active BP and non-BP servers.

Abusive Server Uptime

`MaxiDed` and its merchants shifted the risk of takedown to their customers. They required prepayment, offered no reimbursements, and provided minimal resilience support with considerable attached "abuse fees".

Table 5.4: Server lifespan statistics

Risk Profile	Payment Cycle (days)	Premature Termination (%)	Expired (%)	Extended (%)	Lost Usage (Median # days)	Total (# servers)
No Abuse	91.0	15.69	38.77	45.54	10	4,831
G_1	92.0	18.23	47.39	34.38	23	1,437
G_2	90.0	23.04	52.22	24.74	28	2,834
G_3	61.0	19.59	45.86	34.55	13	3,792
G_4	46.0	15.41	48.39	36.20	3	558
G_5	31.0	19.15	54.73	26.12	6	804

How do customers deal with this risk? In essence: by choosing shorter lease periods for more risky activities. Table 5.4 lists the median lease periods that customers opt for across various risk groups. The more risky the abuse, i.e., the higher the probability of a takedown, the shorter the lease time. The table also provides statistics on the proportions of BP servers that were prematurely terminated due to abuse complaints, proportions of lease expirations, extensions, in addition to the number of usage days that customers lost from termination of their lease. Customers with the most risky activities manage to mitigate the cost of takedown to a median of 6 lost days.

We also see that at most 23% of the BP servers were prematurely taken down. Most BP server ran uninterrupted for their entire lease period. This speaks to the low rate of blacklisting, questioning the

effectiveness of this practices in disincentivizing abuse. An interesting pattern is that customers also abused servers that did not allow abuse. 15% of these servers were also taken down.

Overall 2,656 servers were deactivated prior to the expiry of their lease plan. Another 6,483 active servers were deactivated when they reached their normal expiry term. 5,117 servers remained active beyond their initial lease plan.

Detected Abusive Resources

I next explore a final chokepoint: blocking the BP servers and abusive content hosted on them once they are discovered.

Table 5.5: Statistics on flagged or blocked `MaxiDed` customer resources

	Hosted resources			**Number flagged resource in abuse feed**																
	IPs	**FQDN**	**2LD**	**PHTK**[1]			**APWG**[2]			**SBW**[3]			**GSB**[4]			**DBL**[5]			**CMX**[6]		
Year				(IP)	(FQDN)	(2LD)	(IP)	(FQDN)	(2LD)	(IP)	(FQDN)	(2LD)	(IP)	(FQDN)	(2LD)	(IP)	(FQDN)	(2LD)	(IP)	(FQDN)	(2LD)
2016	985	9,902	3,378	2	1	32	29	45	75	12	10	23	.	.	.	.	.	.	85	185	201
2017	906	15,494	3,573	5	2	18	1	4	23	.	.	.	4	63	71	40	644	696	22	20	51
2018	145	416	280	0	0	2	0	0	5	.	.	.	0	0	4	20	23	22	.	.	.

Notes: (1) Phishing; (2) Phishing; (3, 4) Malware drive-by; (5) SPAM, Malware, Phishing, botnet C&C; (6) Malware and Phishing.
Sources: **PHTK**: Phishtank[124], **APWG**: Anti-Phishing Working Group[123], **SBW**: StopBadware[122], **GSB**: Google Safe Browsing[122], **DBL**: Spamhaus[134], **CMX**: Clean-MX[161].

I triangulated these results by looking directly at several blocklists. I used three years of passive DNS data from Farsight Security's DNSDB [111] to identify domain based resources on `MaxiDed`'s IP addresses: Fully Qualified Domain Names (FQDNs) and 2^{nd}-level-domains (2LDs). Table 5.5 lists the quantities of resources associated with `MaxiDed` from 2016 to 2018. This period corresponds to when `MaxiDed` had the highest number of active servers. I examined the intersection between these resources and those flagged or blocked by several leading industry abuse feeds. The feeds capture a mix of spam, phishing, malware and botnet C&C abuse. Detailed information on these feeds is provided in Table 5.5. The quantities of flagged `MaxiDed` customer resources within each of these abuse feeds are also listed in the table. When no historical feed data was available, I have left the cell empty.

While coverage of blacklists is known to be limited (c. f. [108] or the previous in which I have also discussed their limitation), it is quite disappointing to see the small fraction of the abuse that gets picked up by the feeds. This confirms, with ground truth, the observation in prior work that blacklisting is generally ineffective in disrupting abuse. It also shows the limitations of the metrics that I have produced in previous chapters that they would essentially not be able to flag a provider like `MaxiDed` as an outlier with bad security.

5.10 MARKETPLACE FINANCES

Disruption of BPH is also determined by how profitable the business is. Lower margins mean that the provider is more vulnerable to raised operating costs in the supply chain. In this section, I analyze MaxiDed's revenue, costs and profits. To get a sense of the company as a whole, I include both BP and non-BP services.

(Revenue.) From the 23 different payment instruments employed by MaxiDed, most of its revenue was received via WebMoney payments (1,493,876 USD) followed by direct BitCoin payments (1,324,449 USD, MaxiDed itself logged these in USD). Around 577,118 USD was received through the remaining payment instruments. The total amount of revenue from 2011 up to May 2018, adds up to 3.4M USD.

(Operating Costs.) We have no data on personnel cost at MaxiDed. Here, I analyze the outgoing payments to merchants, upstreams and outstanding debts recorded in the database.

i) *Payments to Merchants.* A main component of MaxiDed's cost structure consists of payments to merchants. Merchant payments were exclusively deposited on WebMoney and Epayments wallets. After MaxiDed took their 20% fee, the remaining 80% went to the merchants. Analyzing outgoing MaxiDed payments show 11 of the 14 operating merchants to have received payments, adding up to 1,588,810 USD.

Figure 5.14 illustrates the distribution of payments made to each merchant. The two largest suppliers of server packages, mc_4 and mc_{14}, received the bulk of the earnings. Most of the merchants were completely unsuccessful. The lowest earners, combined, generated less than 190K USD over all years.

ii) *Payments to Upstreams.* We cannot see the payments of third-party merchants to their upstreams, only the payments where MaxiDed is itself a merchant on the platform (mc_0). The data shows that mc_0 payments to their upstreams add up to 1,526,015 USD, paid via WebMoney and PayPal. Note that 99% of these payments were not for BP servers, as those were almost exclusively provided by the third-party merchants.

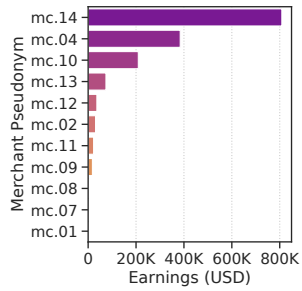


Figure 5.14: Payments to merchants.

iii) *Debtors.* The final component of MaxiDed's costs structure is that of outstanding debts due from its customers. The operators have vigilantly banned customers with outstanding debts. One customer was the exception to this rule. Actually, this was not a real customer, but a customer account through which MaxiDed operators themselves purchased servers from merchants on their platform. These were used to host DepFile, their large file-sharing platform

side-business. This customer entity accumulated debts amounting to 399,123 USD.

(Profits.) Table 5.6 details MaxiDed's yearly finances, alongside finances of their side business DepFile. Despite the common understanding of BPH services being lucrative, we clearly observe MaxiDed's earnings to be modest and declining. In total, over seven years, MaxiDed made just over 280K USD in profit. If we take out the debt incurred for the DepFile side-business (399,123 + 280,618), then the profit would have been 679,741 USD. This is still an underwhelming figure for 7 years of operating a BPH platform. Recall that the cost of personnel, office space, and equipment also has to be taken from this amount. These combined costs would have to be substantially lower than 100K USD per year to leave even a tiny profit on the balance sheet. Also note that there is large asymmetry between these profits and overall costs of cybercrime also noted by Anderson et al., were in fact these profits are dwarfed in comparison to the negative externalities that are caused by allowing abusive content to be hosted [71].

Table 5.6: Overview of MaxiDed's yearly finances along side that of DepFile

	MaxiDed			DepFile			
Year	**Revenue**	**Costs**	**Prof$_{mx}$**	**Revenue**	**Costs**	**Prof$_{dp}$**	**(Σ Prof.$_i$)**
2011	79,987	1,312	78,675	.	.	.	78,675
2012	345,213	72,418	272,794	.	.	.	272,794
2013	458,028	17,9761	278,266	334,540	248,307	86,233	364,499
2014	419,739	328,757	90,981	1,646,568	712,442	934,125	1,025,106
2015	615,046	570,895	44,150	2,205,687	1,396,820	808,867	853,017
2016	733,151	726,040	7,111	3,153,553	2,188,634	964,919	972,030
2017	566,471	872,520	-306,048	3,998,244	2,841,322	1,156,922	850,874
2018	177,806	363,118	-185,312	1,547,078	1,129,586	417,492	232,180
Total	3,395,444	3,114,825	**280,618**	12,885,673	8,517,113	**4,368,560**	**4,649,178**

Note: (*mx*: MaxiDed) (*dp*: DepFile)

The side-business DepFile, on the other hand, appears to have generated much better margins. We could even speculate that MaxiDed was more valuable to its owners as a way to acquire cheap and risk-free server infrastructure than as its own profit model, were in fact the better profit margins are to be gained in other forms of cybercrime.

5.11 RELATED WORK

(Underground Ecosystems.) Several ecosystems and marketplaces of a malicious nature have been studied in the literature via captured datasets. Stone-Gross et al. analyzed credential stealing malware [162]

and spam botnets [26] by taking over part of the botnet infrastructure to understand their inner workings. Wang et al. studied SEO campaigns to sell counterfeit luxury goods and the effectiveness of various interventions to combat such activities [64]. Alrwais et al.[163] investigate illicit activities in the domain parking industry by interacting with the services to collect ground truth data. Christin [31] analyzed the Silk Road marketplace by running daily crawls of its webservices for 6 months to understand merchants, customers, and what was being sold. A followup study by Soska and Christin [164] examined 16 anonymous market places also by periodically crawling their webservices and found that marketplace takedowns may be less effective than pursuing key merchants that may migrate to others. Another followup study by Wegberg et al. [32] augments previous studies by examining evidence for commoditization of entire cybercrime value-chains in underground marketplaces and finds that only niche value-chain components are on offer.

Datasets on the underground can also be leaked by criminal competitors. McCoy et al. used leaked databases of three affiliate programs to study pharmaceutical affiliate programs [29]. More recently, Brunt et al.[165] analyzed data from a DDoS-for-hire service and found that disrupting their regulated payment channel reduced their profitability but that they were still profitable by switching to unregulated cryptocurrency payments. Hao et al. [156] analyzed a combination of leaked and legally seized data to understand the ecosystem for monetizing stolen credit cards. The dataset used in this study resulted from the aftermath of the legal takedown of the BPH provider `MaxiDed`. To the best of my knowledge, there has been no prior academic work on BPH using such ground-truth data. This study presented in this chapter uniquely provides a comprehensive picture of the supply, demand and finances of the entire BPH operation.

(Bulletproof hosting.) Earlier efforts on detecting BPH have relied heavily on identifying autonomous systems. Fire [36] was one of the first systems for detecting BP ASes by temporally and spatially aggregating information from multiple blacklists in order to detect elevated concentrations of persistent abuse within an AS's IP blocks. Shue et al. [92] noted that BP ASes often fast-flux their BGP routing information to evade detection. ASwatch [93] leveraged fast-fluxing BGP routing as strong indicator of a BP AS to build a classifier and detect BP ASes before they appear on blacklists. My own studies discussed in previous chapters have developed security metrics to compare concentrations of abuse on various hosting networks and to identify negligent providers that may be suspected of operating BPH services [100, 101], while Tajalizadehkhoob et al. developed techniques to analyze abuse concentration on the hosting market as a whole by identifying providers from their WHOIS information rather than BGP data [44]. BPH however, has evolved

over time. Alrwais, et al.[37] studied a recent approach of BPH abusing legitimate hosting providers through reseller packages to provide a more agile BP infrastructure. This chapter complements this work by providing a unique perspective into to the ecosystem of BPH. Based on my analysis, we can better reason about which mitigation techniques might be effective and which are likely ineffective for undermining modern agile BPH marketplaces.

5.12 LIMITATIONS AND FUTURE WORK

In comparison to other underground marketplaces studied previously (cf. [32, 164]), `MaxiDed` may be seen as a specialized marketplace for provisioning BP servers. While comparisons with other underground markets may be drawn, direct comparisons are difficult due to differences in how `MaxiDed`'s marketplace operated. For example its customers were not aware that merchants were involved in supplying the marketplace with resources. This also explains why in comparison no reputation mechanisms were in place for customers to differentiate packages based on their quality (or differentiate good/bad merchants).

Despite such differences, I do still observe patterns similar to what other studies of criminal endeavors have reported. For example, I have observed a concentrated supply pattern around a handful of merchants in `MaxiDed`'s case, which is a similar to what other studies of underground market places have observed ([32, 164]). I have also observed demand to gravitate towards the resources supplied by successful merchants. The number of successful merchants being limited, also agrees with studies of other criminal operations, e.g. in studying spam botmasters and their operations [26].

Given that this study has focused on an in-depth analysis of the anatomy and economics of `MaxiDed`, future work may draw more systematic comparisons to better understand the implications of what has been reported here. Furthermore, `MaxiDed`'s prominence within the ecosystem has also not been systematically explored in my study, albeit the limited comparisons with other BPH providers in addition to anecdotal evidence [143, 155] suggest that `MaxiDed` may be reasonably considered as a major provider within the ecosystem. Nevertheless, some of my findings, particularly those relating to the economics and profitability of BPH services may require further research to better understand the BPH ecosystem as a whole.

5.13 DISCUSSION AND IMPLICATIONS

(Discussion.) I found `MaxiDed` to have developed a new agile model in response to detection and disruption strategies. Its operations had ma-

tured to the point of a new innovation, namely operating a marketplace-like platform for selling BPH services. This model transfers the risks of acquiring the BP server infrastructure from upstream providers to merchants. `MaxiDed` 's main role was to take on the risks of acquiring customers, communicating with them and processing their payments. The 14 merchants on the platform (over)-supplied the market with more than 50*K* different server packages, many of which expired without being purchased. They abused a total of set 394 different upstream providers, thus allowing merchants to spread out and rotate abuse across many different legitimate networks.

I see some concentration in this supply chain, with 15 upstreams providing infrastructure for over 50% of the BP servers sold. Most of these upstream resources are not shown to be delegated in WHOIS, drastically curtailing the effectiveness of the most recent detection approaches. Another point of concentration is in the merchant pool: two merchants offered 89% of all BP servers and made 94% of the BP packages sales. Most other `MaxiDed` merchants failed to generate any meaningful sales. The platform deployed 23 different instruments to transact with customers over various periods. Revenue was initially largely processed by one payment settlement system: `WebMoney`. We also saw an increased volume of `BitCoin` payments and the adoption of other cryptocurrencies in response to disruptions in other instruments, such as `PayPal`. A lack of product differentiation on the market is likely to have created a fierce price competition across the merchants which in turn has led a great proportion of merchants to fail. This competition also decreases the profits of not only the merchants, but also of `MaxiDed` itself. Its profits, over seven years, amounted to a mere 280*K* USD (or 680*K* USD if we ignore cross subsidies to their other business, `DepFile`). The actual profits are even lower, as this amount also has to cover the cost of personnel, office space and equipment, on which I had no data.

(Implications.) Bullet-Proof Hosting (BPH) companies remain a difficult problem as their operators adapt to evade detection and disruption. Prior work in this area has largely relied on external measurements and generally lacks ground-truth data on the internal operations of such providers. Recent detection techniques rely on certain assumptions, namely that agile BPH operates under reseller relationships, and that upstream providers accurately reflect such relationships in their WHOIS information. I found `MaxiDed` to deviate from both assumptions, thus rendering detection less effective.

Prior BPH instances were mainly disrupted by pressuring upstream providers to sever ties with downstream BPH providers. Given the number of available substitute upstream providers of `MaxiDed`, this is unlikely to be an effective chokepoint. Drawing parallels with other underground markets suggest that, other than taking down the platform itself, disruption may also be achieved by pressuring other chokepoints:

merchants, revenue and demand. MaxiDed's dominant merchants would have been a viable chokepoint, yet, identifying them most likely required internal operational knowledge as their existence and identities were not externally visible. As for disrupting payment channels, the transition to mostly unregulated cryptocurrencies payments suggest that this is no longer a straightforward option. Surprisingly, MaxiDed's low profits indicate that an increase in transaction or operating costs may be viable a pressure point to disrupt revenue and demand. Future work could explore how to raise these costs. Being aware of the threat of criminal prosecution might, ironically, be one way.

The final remaining pressure point would be to take down the platform. Such takedowns however are hard to replicate, let alone scale. That being said, MaxiDed explicitly marketed bullet proof services on the clear web. Even in cases when criminal prosecution itself is not feasible, if the threat can be made plausible, it might force the company to operate within higher op sec requirements, raising the cost of doing business. This suggests that what appears the more difficult strategy might actually be the best option in light of the supply chain becoming even more agile and evasive. My hope is that by further studying and understanding these emerging agile BPH services we can inform new and potentially more effective directions for mitigating this threat. To orient future work in this area, researchers might be better off deprecating the increasingly misleading metaphor of "bullet-proof" hosting in favor of a term like "agile abuse enablers".

5.14 ADDITIONAL MATERIAL

A - Customer Preference Elicitation

Customer : Some servers don't specify what is allowed. does this mean everything is OK?

Provider: What are you looking for?

Customer : I'm looking for malware, spam and botnet C2 hosting , VPS or physical server are both fine

Provider: We allow this here for example ... [provides link to server package configurator]

Customer : That says xrumer, warez, adult, ...not what I asked for

Provider: We don't mention what you want on the public list

Customer : Can you send me a large private list to choose from?

Provider: [provides link to dedicated servers located in a country]

Provider: Dedicated server prices are above 100

Customer : All of these are in one country, anything in US or EU?

Provider: [provides several links to other server package configs]

Figure 5.15: Chat excerpt illustrating customer preference elicitation.

Figure 5.15 illustrates an excerpt of a live chat (edited for readability) conducted by one of the authors with `MaxiDed` operators prior to its takedown. It shows the process of preference elicitation by `MaxiDed` operators.

The conversation was conducted using the live-chat functionality on their webshop. It demonstrates that `MaxiDed` operators may have also allowed other forms of abuse which they did not publicly mention on their webshop along side the various BP server packages that the platform advertised.

B - Geographical distribution of Customer Servers

In analyzing `MaxiDed`'s platform, I also examined where its customer servers were located. I used Maxmind's commercial historical geolocation data for this purpose. This data is available on a weekly basis. For each customer server I first found the closest matching Maxmind IP geolocation database with the timespan during which the server was active. I then determined were each server was located based on its IP address and Maxmind's datasets. Figure 5.16 plots the top-20 locations for `MaxiDed`'s customer servers.

I found that the majority of the BP servers geolocated to Moldova followed by Russia, the US, Ukraine, the Netherlands and a long tail of other countries. Figure 5.16 also displays the number of non-BP servers in each of these top-20 locations. I observed that the Netherlands in particular hosted a substantial number of the non-BP servers.

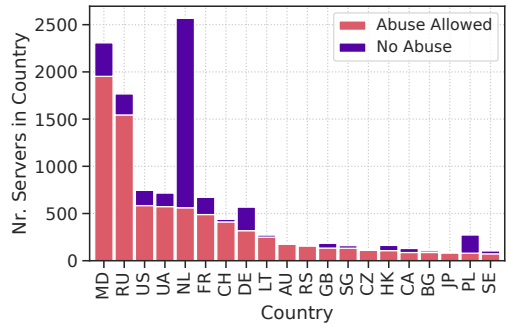


Figure 5.16: Top-20 locations for `MaxiDed` customer servers

6

DDOS VICTIMS AND THE EXTERNALITIES OF SECURITY NEGLIGENCE

This final content chapter is based on a published study of mine [104], and employs some of the methodological contributions of my previous chapters towards examining how security failures in the hosting market negatively affect others. It examines the negative externalities of so-called 'booter' websites, which package and sell the ability to kick arbitrary targets offline by launching Distributed Denial of Service (DDoS) attacks at the click of a button. In essence they provision and sell DDoS-as-a-Service at often low prices to anyone that is willing to pay for it. As part of their service, booter websites typically also amplify attacks in comparison to more traditional DDoS attacks. That is by employing so-called amplification techniques booter websites boost attack volumes by playing on the asymmetry between the volume of generated traffic which is directed at the victim versus the actual amount of traffic required to generate the attack. The empirical insights gained in this study demonstrate the extent of harm that malicious content may create if not adequately addressed, particularly by hosting providers that host such booter websites and are in a unique position to take them offline.

Within this area, a lot of research has been devoted to understanding the technical properties of DDoS attacks, particularly how amplified attacks may be generated, and the emergence of the DDoS-as-a-service economy, especially these so-called booters. Much less however, is known about the consequences for victimization patterns. As such this final chapter profiles victims via data from amplification DDoS honeypots to provide a broader understanding of how security negligence in the hosting market plays out.

I conclude this chapter by also reflecting on the implications of my findings for the wider trend of commoditization in cybercrime.

6.1 INTRODUCTION

While Distributed Denial-of-Service (DDoS) attacks have been around for a long time, the use of amplification techniques has transformed the criminal ecosystem. These techniques now make up the bulk of the observed attack traffic [149, 166]. This shift is intimately related to another trend: the rise of DDoS-as-a-service, also known as *booters*. Booters are a clear example of the so-called commoditization of cybercrime [24]: criminal service providers bundling all the resources and

tools needed for an attack and offering them in an accessible way as a commodity service to anyone willing to pay.

Several in-depth studies have illuminated the supply side of the market for DDoS: the technical resources and techniques deployed by the criminal service providers [166, 167, 168]. We have also learned quite a bit about the economics of booters from publicly-leaked dumps of several operational databases containing information about revenue and customers [169, 170, 171].

What is much less understood, however, is how the abundance and affordability of DDoS-as-a-service has impacted victimization patterns. Who is bearing the brunt of the lowered barriers for DDoS attacks? Existing studies have revealed some basic distributions of victims across countries, Regional Internet Registries (RIRs) and Autonomous Systems (ASes). They have pointed to end hosts, gaming servers and hosting providers [149], but they lack a more in-depth investigation and explanation of victimization patterns.

This chapter addresses this knowledge gap and profiles the affected networks and victims. It uses a dataset of 1,115,795 victim IP addresses captured over two years (2014-2015) via several amplifier-honeypots [166]. From the IP addresses, I have inferred certain properties of the victims and identified the factors determining their distributions across networks and countries.

Since the existing work on amplifiers and booters has not focused on the victims, the public understanding of them has been shaped by anecdotal news articles and by industry reports compiled by DDoS mitigation providers. The former focus on the more news-worthy cases, i.e., the attacks against high profile targets. The latter are biased towards their own customer base, i.e., enterprises purchasing DDoS protection services, as that is where the data is being collected. As I demonstrate in this chapter, neither provide a good understanding of the ecosystem of commoditized DDoS attacks.

The main contributions of this chapter are as follows:

- In this chapter I show that the bulk of the victims (62%) are users in access networks, rather than in hosting networks (26%). Only a small fraction resides in enterprise networks;
- I demonstrate that the victimization rate in access networks is highly proportional to the number of broadband subscribers in those networks, suggesting that the commoditization of attacks has led to a democratization of victims;
- I find that certain countries have a significantly higher number of victims per subscriber. This rate is weakly related to institutional factors such as Information and Communication Technology (ICT) development, suggesting geographical network effects among attackers and victims increasing the uptake of DDoS-as-a-service;

- I demonstrate that victimization in hosting networks is proportional to the number of IP addresses and hosted domains, and also influenced by the popularity of the hosted content.
- Where I was able to specifically identify webhosting victims, I find that they have barely any enterprises among them or other valuable targets. The largest victim group are gaming-related sites, most notably around Minecraft, suggesting that the commoditization of DDoS facilitates crime that is mostly not profit driven.

In what follows I first present some background (Section 6.2) and the data collection method (Section 6.3), I then discuss the distribution of victim IP addresses over access, hosting and other networks (Section 6.4). Next I delve deeper into victimization patterns in access networks (Section 6.5) and hosting networks (Section 6.6). I then briefly explore whether attack duration is different across victim populations (Section 6.7). After comparing my findings to related work, I summarize my conclusions on the consequences of DDoS-as-a-service and discuss the implications for the wider issue of the commoditization of cybercrime.

6.2 BACKGROUND

DDoS attacks have been associated with a range of motives. They can be profit-driven – as in the case of extortion, disrupting competitors, or using it as a smoke screen for committing financial fraud – or motivated by other objectives, such as political protest, harassment, or gaining advantage in online gaming [24, 149].

Amplification DDoS attacks now make up a considerable fraction of network-layer DDoS incidents [172, 173, 174]. Attackers send requests to amplifiers – a.k.a. reflectors – and spoof the source IP address, so that the amplifiers' responses are directed to the victim. A whole range of protocols can be abused for amplification and millions of machines run these protocols which enables such attacks [139].

Most of the amplification attacks stem from booter services [166, 170]. The price for purchasing an amplified DDoS attack can be as low as $1, as the analysis of some leaked booter databases demonstrates [170, 175]. A purchase from a booter would typically entail access to the service for a limited amount of time, tied to different pricing tiers. Most attacks are very short, less than 10 minutes [170].

On the customer side of booter services, leaked databases have shown that most customers of DDoS-as-a-service use it only once to attack a single target [170] and only a small fraction of them hide their tracks via `Tor` or `VPN`. This might indicate that their technical skills are limited or that they do not perceive a need to hide. The users that do hide

their tracks, tend to return for more and also tend to launch more attacks [169]. The databases have also revealed that gamers make up a specific and important customer group [169]. On the victim side, booter databases contain the targeted IP addresses or URLs, but these sets are limited in scope and volume. The top 100 most attacked sites were mostly game servers and game forums [169].

Besides booter databases, NTP amplification attacks allow victim IPs to be retrieved from the NTP servers. From this data, Czyz et al. point to end hosts and gaming servers to be common victims [149]. Amplification honeypots have also collected victim IP addresses [166]. They have only been superficially analyzed, in terms of the distribution over countries and IP address space. The U.S., China and France were the most attacked countries. In this chapter, I significantly extend the analysis of honeypot data.

The only other systematic source of information comes from industry reports by DDoS mitigation providers. Akamai points to gaming, software and the financial industry as the major victims [172], with a small fraction of victims belonging to the telecom industry. Other reports suggest hosting as major victims [176]. These industry reports have specific limitations and biases, which I will return to in Section 6.4.

6.3 HONEYPOT DATA

The victim data used in this study was gathered via a set of amplifier honeypots – dubbed AmpPots [166] – which are still deployed to monitor attacks. My data pertains to two years of data over 2014-2015 from these AmpPots. They run services that are known to be misused for amplification attacks: QotD (17/udp), CharGen (19/udp), DNS (53/udp), NTP (123/udp), SNMP (161/udp) and SSDP (1900/udp). Each AmpPot uses real server software (in 'proxy' mode) to provide the aforementioned services except for SSDP in which an emulated script is used instead. The responses of AmpPots are filtered in order to prevent from contributing to actual attacks. More details of AmpPot can be found in prior work [166] which I refer the reader to.

In total 8 AmpPots were deployed on the Internet during the measurement period of 2014-2015. Table 6.1 shows a summary of the operational timeline and supported protocols of these devices. At the start of the measurement period (2014-01-01), two AmpPots were operational and initially only supported the CharGen and DNS protocols. With a sustained effort to monitor more amplification attacks, more devices were gradually added with support for additional abused protocols. At the end of the measurement period (2015-12-31) the deployed AmpPots collectively monitored 6 services except for H02 which only supports DNS. All AmpPots are located at ISPs in Japan and their IP addresses are

Table 6.1: Overview of deployed AmpPots.

AmpPot ID	Deployed on	IP Changes	Notes
Ho1	2012-10-07	19	added QOTD, NTP, SNMP, SSDP on 2014-09-25. Discontinued on 2015-10-09
Ho2	2013-05-13	25	only DNS supported
Ho3	2014-05-13	9	added SNMP support on 2014-09-17 and SSDP on 2014-10-03 *
Ho4	2014-05-13	10	added SNMP, SSDP support on 2014-09-17 *
Ho5	2014-05-10	4	added SNMP, SSDP support on 2014-10-18 *
Ho6	2014-05-10	6	added SNMP, SSDP support on 2014-10-18 *
Ho7	2014-05-10	8	added SNMP, SSDP support on 2014-10-18 *
Ho8	2015-11-09	0	_ **
Note:*			Deployed with QOTD, CharGen, DNS and NTP support
Note:**			Deployed with support for all protocols

dynamically assigned. Depending on the ISP, the IP addresses changed every 5-30 weeks, on average.

AmpPots observe not only amplification attacks, but also scans from researchers or attackers who search for vulnerable devices. To separate actual attacks from scans, attacks are defined as a series of at least 100 consecutive query packets that a single host sent to an AmpPot, where consecutive means that there was no gap of more than 600 seconds between two packets. This definition is in concord with the one used in [166]. I did, however, reduce the gap from 3600 seconds to 600 seconds, in order to analyze attack duration with a more fine-grained approach.

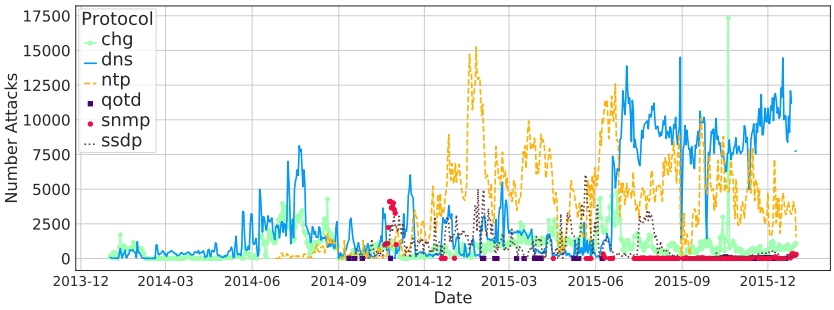


Figure 6.1: Number of amplification attacks per protocol observed in AmpPot data over the course of 2014 and 2015

Collectively, the AmpPots have monitored 1,115,795 unique victim IP addresses from 92 countries and 15,044 unique victim ASes. Figure 6.1 shows the number of attacks per protocol during 2014 and 2015. As the figure demonstrates, the total number of attacks has increased over time and protocols like `DNS`, `NTP` and `SSDP` have been used more often to launch amplification attacks. During the measurement period, the AmpPots have monitored 5,726,150 amplification DDoS attacks in

total: DNS (41.26%), NTP (38.73%), CharGen (11.32%), SSDP (8.01%), SNMP (0.65%), and QotD (0.01%).

6.4 VICTIMS OF AMPLIFICATION ATTACKS

Given the amplification attack data the first question I pursue is: *In which type of networks are victims concentrated?*

To avoid confusion, I first define the main concepts. The term *attack* has been defined and operationalized in the previous section. I use the term *target* to refer to the entity (or entities) that the attacker intended to affect. This may be a person, organization, service or machine. Since the data consists of IP addresses, the attacker's intention is not directly observable. For this reason, I use the term *victim* to refer to the targeted IP addresses and the hosts residing there. As DDoS attacks are also a cost to the networks in which the victims reside, I refer to the Autonomous System (AS) that routes the traffic for the victims as *victim AS* or *victim network*. To answer my question I looked up the ASes of the victims and categorized them into three groups: *broadband ISP*s, *hosting* providers, and *other* networks.

To reliably identify the broadband ISPs, I utilize a previously developed mapping that identifies the ASes of broadband ISPs in 82 countries, that has been previously used to study botnet mitigation in broadband ISPs [82]. The mapping accurately distinguishes between and provides labels for Autonomous System Numbers (ASNs) which have been manually mapped to broadband ISPs, hosting, governmental, mobile ISP, educational and other types of networks. In total, the mapping contains 2,050 labeled autonomous systems. The mapping is organized around ground truth data in the form of a highly accurate commercial database; *TeleGeography Globalcomms* [177], containing the broadband subscriber numbers of 211 countries. Compared to machine learning approaches that map AS types [178], this mapping is more accurate since it manually identifies access networks, and the completeness of the mapping is verified with the Telegeography database.

To identify *hosting* providers, I take all the non-broadband ASes in the data and apply a simple heuristic to them similar to the ones that I previously used in Chapter 2, Chapter 3 and Chapter 4. First, I count the number of unique 2^{nd}-level-domains (2LDs) hosted within the ASes. For this I used all observed domains in 2014 and 2015 in DNSDB, a large passive DNS (pDNS) database operated by Farsight Security [111]. DNSDB is sourced from more than 100 sensors located around the world, in addition to authoritative DNS data from various top-level domain (TLD) zone operators. To illustrate: in 2015 DNSDB observed 287M unique 2LDs, which map to 69M distinct IP addresses.

I use the accurate AS labels mentioned above to identify a threshold for the number of hosted domains per AS that most accurately separates

the ASes labeled as hosting from other types of ASes which may also host domains. This approach does mean that Content Distribuion Networks (CDNs) and others networks like `Cloudflare` also get categorized as hosting. Based on a constructed Receiver Operating Characteristic (ROC) curve I identify this threshold to be 2,700 2LDs. Therefore I define as hosting any AS that has not been previously identified as a broadband ISP which hosts more than 2,700 2LDs. This corresponds to a false-positive/true-positive rate of 0.17/0.74. This accuracy is far from perfect, but better than available alternatives. I compared it to machine learning approaches, such as `CAIDA`'s classification of ASes [178]. Using `CAIDA`'s '`Content`' label as an alternative for classifying the hosting providers results in a 0.04/0.32 false-positive/true-positive rate of classification. This classification has a better false-positive rate, but this comes at the cost of a highly reduced true-positive rate in comparison to my classification. Alternative methods for identifying hosting providers have also been explored in [44] and discussed in Chapter 3 and Chapter 4. They are not directly comparable due to their organizational level classification rather than AS level but I have adopted the same heuristic technique for classification.

Finally, all ASes that have not been classified as broadband ISP or hosting are labeled as *other*. Manual inspection show that this group contains governmental and educational networks, mobile and cloud providers, enterprises and more.

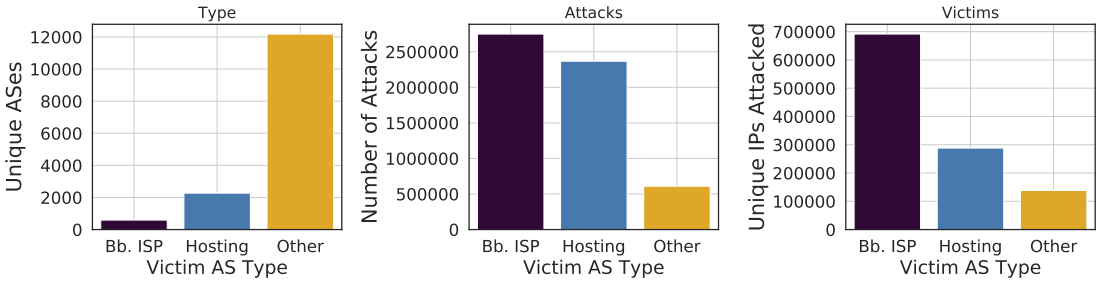


Figure 6.2: Spread of unique victim networks (left), attacks (middle) and unique victim IPs (right) observed in the data over the various types of networks that have been identified

Having constructed the network classification, I can now examine the distribution of victims over these networks. Figure 6.2 plots the results.

It clearly shows that the majority of attacks and victim IPs are located in broadband ISPs, even though they only constitute a small fraction of all ASes that have been attacked. More precisely, 48% of the attacks and 62% of the victims are in access networks. In total, I observe victim IPs from 92 countries in the attack data. We have detailed information

on ISPs from 77 of these 92 countries. All identified ISPs in these 77 countries receive attacks, except for 5 countries (GB, US, JO, KE, LV) where at most 2 smaller ISPs are missing from the attack data. This suggests that the whole global broadband market is victimized by these attacks.

The second largest category is hosting: 41% of attacks and 26% of victims. The remaining victim networks constitute only a small fraction of the attacks and victims (11% and 12%, respectively).

This distribution of victims across broadband and hosting networks is different from earlier work by Czyz et al. [149]. They observed that the top 10 most targeted networks consisted of eight hosting providers and two telecom companies and that access nodes made up around half of all victims. They did observe already a trend that the portion of victims in access networks was increasing, which seems to have continued after their measurement period. My analysis of the `UDP` ports used for the attacks largely agrees with that of [149]. The most frequently attacked `UDP` ports by a large margin include ports 80 and 8080, that are more likely to be open and accessible through firewalls. Other application specific ports are also targeted such as (7000) for BitTorrent trackers and CORBA management agent (1050).

I have triangulated our results with `CAIDA`'s mapping of ASes [178], which classifies them as '`Content`', '`Enterprise`' or '`Transit/Access`'. While these category labels are different from my classification of networks, which means we cannot directly compare the exact distributions, the CAIDA mapping also locates most victims in `Transit/Access` networks, followed by `Content` and `Enterprise`. This is consistent with my findings.

Networks are not homogeneous, of course. Broadband networks, for example, can also contain hosting services. To probe deeper into the AS-level pattern, I take a closer look at the IP addresses of victims in access and hosting networks. I checked whether the addresses were associated with any domains in pDNS data. Domains are used for a variety of hosting services; websites, but also for gaming servers, email servers, basically for any service where a human readable name is more convenient than an IP address. The pDNS data found that 95% of the victims in broadband networks have never been associated with any domains in 2014 and 2015. This suggests that the bulk of the victims in these networks are access nodes. The remaining 5% host on average 20.8 domains per IP address (The median domain count is 1 and 75% of these victims host 3 or less domains).

Since this categorization is dependent on the coverage of the pDNS data, I have cross-checked our domain data with the *Bing.com* search engine. I took a random sample of 1,000 broadband victim IP addresses and queried Bing ('IP:<x.x.x.x >') to see if any domains were associated with it. For 9% of the cases, BING reports observing domains where

our pDNS data did not observe any. The opposite was true in 2% of the cases. This suggests that the pDNS data gives a reasonably accurate picture.

In hosting networks, I found that 46.6% of the victim IPs have been associated with domains. This confirms earlier work that webhosting is just one among many targets. Figure 6.3 summarizes the breakdown of the victim types and the subsets which I analyze in more detail in subsequent sections.

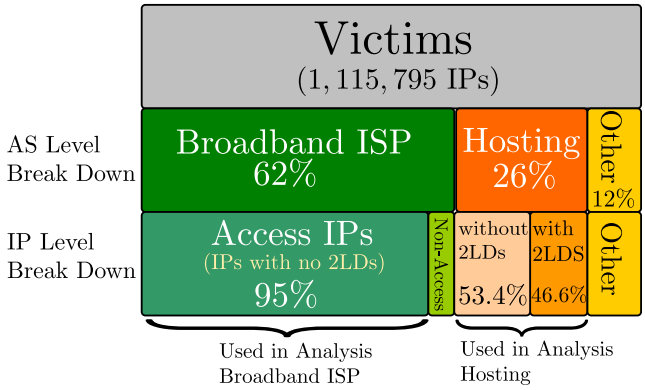


Figure 6.3: Breakdown of DDoS attack victims observed in the AMPPOT data

Note that these results substantially differ from the victimization analysis in the reports of DDoS mitigation providers. Essentially there are two types of industry reports: based on traffic data or based on customer surveys. An example of the former is Akamai's State of the Internet report [179]. It identifies the gaming industry as the largest victim of DDoS attacks with 54% of the attacks, followed by the software and technology industry (23%) and financial industry (7%). Only 4% of attacks map to the Internet and Telecom industry. Another type of industry report is based on surveys among customers of DDoS mitigation providers. A recent example is Arbor Networks' WISR [173], which surveys 287 different organizations of which 24% are ISPs and 5% hosting providers. Other industry reports [176] point to hosting as the main victim however, this could be due to a focus on botnet-assisted DDoS attacks rather than amplification attacks.

The mismatch between these reports and my findings is evident. I would argue that when it comes to observing victimization, the industry analyses are more biased than the honeypot data. Industry data is typically collected in the networks of the customers of the DDoS mitigation providers. It is unlikely that users in retail broadband networks are purchasing these kinds of services and thus those victims are severely under-counted by the industry reports. The amplifier data is much less biased towards certain types of victims. This strength does

come at the cost of a weakness: missing attacks that are not amplifier-based. Earlier work suggests this is not a significant issue. Czyz et al. compared the data captured by observing NTP amplifiers against industry measurements and victim network data and they found that the patterns observed in the amplifier data were consistent with the industry measurements [149].

The contrast between my findings and industry reports are more than measurement issues. They have significant theoretical implications for our understanding the DDoS ecosystem, a point to which I will return later in the chapter. But first I turn to a more in-depth look at the victimization patterns in broadband ISPs and hosting networks.

6.5 VICTIMS IN BROADBAND PROVIDERS

I have now established that the majority of victims reside in broadband provider networks and that the majority of these victims are access nodes. In other words, home routers are typically the most affected devices. It suggests that the actual target is a regular home user behind that router. This brings us to the next question: *How are victims distributed over broadband networks?*

A simple count of unique victim IP addresses over the whole measurement period, does not give us a decent metric of victimization rates per ISP because of `DHCP` churn. ISPs re-assign IP addresses to their users at varying rates, where high rates lead to significant over-estimation of the number of victims. One method to reduce the effect of churn is to use short measurement windows [82, 117]. For this reason, I count the unique number of IP addresses seen for each day and then average those daily counts to get to victimization rates over larger time frames. This results in a more accurate representation of the relative victimization rate per ISP.

In Figure 6.4, I have plotted the average daily number of victims against the number of subscribers of those ISPs. The subscriber data is drawn from the `TeleGeography` database discussed in the previous section [177]. The database provides accurate worldwide subscriber numbers for ISPs from 77 countries that appear in our attack data. It provides a more precise proxy for the number of users in a network than technical network properties, like the number of advertised IP addresses, can provide.

As we can see, victimization rates differ by several orders of magnitude across ISPs, but these differences are highly correlated with the size of the customer base: $R^2 = 0.60$. As an aside, the correlation with the number of IP addresses advertised by each ISP also shows a strong linear relation, though a bit weaker ($R^2 = 0.56$).

In other words, the number of users is a strong predictor for the number of observed victims. This is consistent with the earlier speculation

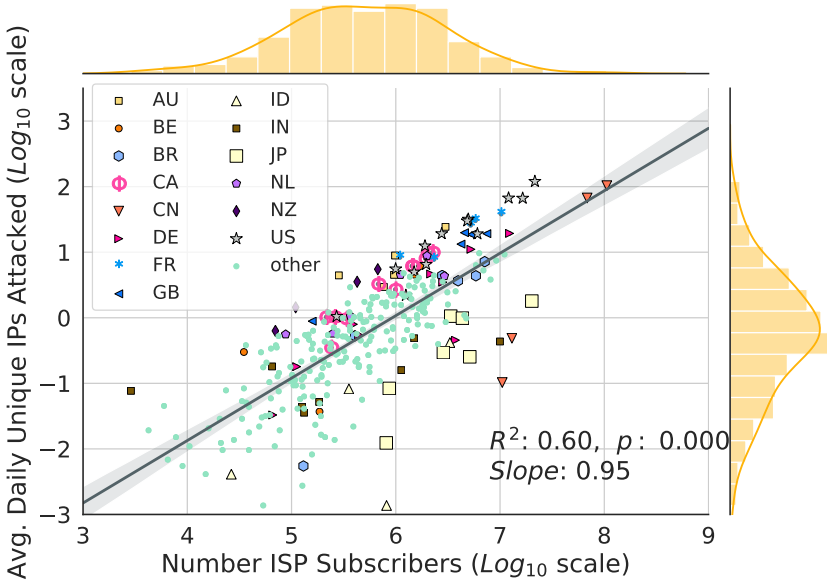


Figure 6.4: Correlation between number of victims in broadband networks and number of ISP subscribers. Histograms along the axes depict the distribution of data points along each axis indicating their normal distribution.

that it is individual users that are being attacked, rather than services or devices. It also means that, to some extent, victimization rates are uniform across ISPs. Whatever motivations attackers may have, it seems they select targets somewhat evenly across broadband networks.

Notwithstanding the effect of the size of the subscriber base, as captured by the regression line, the figure also clearly shows that there is significant variation around that line. That raises a new question: *why do some ISPs have disproportionately more or fewer victims?* We can use the victim rates of ISPs (i.e., the daily average number of victim IP addresses divided by the number of ISP subscribers) to further explain the variance among them. From the size-corrected victim rates we can see that there are several orders of magnitude differences among the most and least attacked ISPs. How can these differences be explained?

In Figure 6.4, I have color coded ISPs by the country in which they operate. To better highlight between and within country relations, Figure 6.5 plots the distribution of ISP victims per subscriber in relation to the country in which they operate. Two things become apparent. First, in many countries, ISP victimization rates are remarkably clustered, compared to the overall variance across countries. Second, ISPs in some countries are victimized less, according to my metrics. In other words, there seem to be country-level effects at work, in addition to network- and user-level effects. The plot shows that ISPs in New Zealand, Aus-

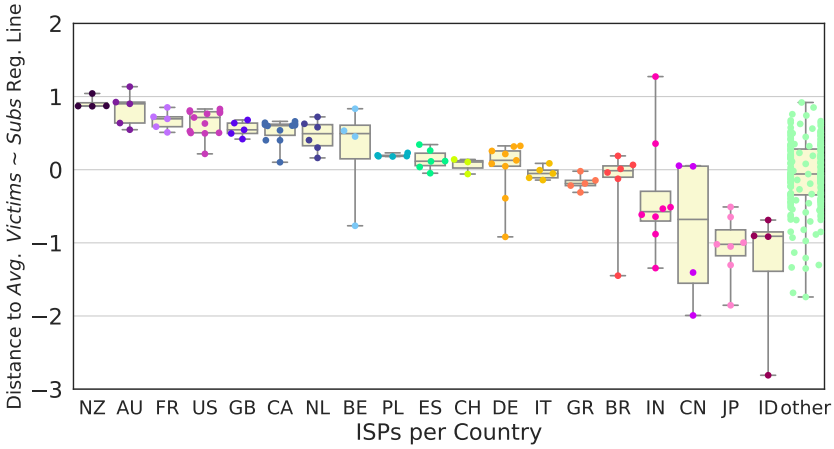


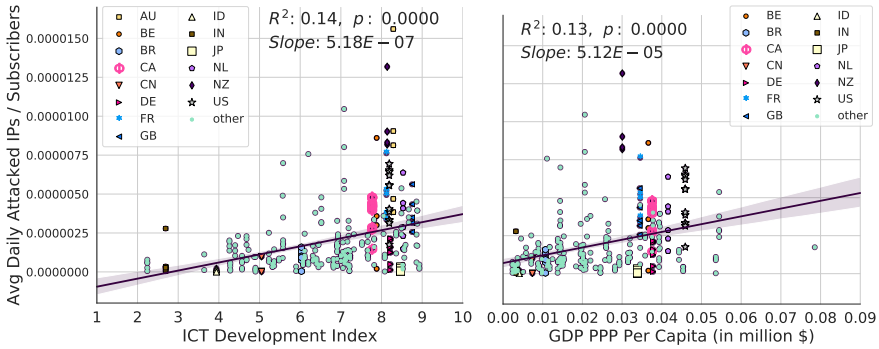
Figure 6.5: Between and within country differences among victimization in different ISPs

tralia, U.S., U.K. and France have disproportionately more victims, while ISPs within countries such as China, Japan and Indonesia have disproportionately fewer. It is important to note that almost all ISPs in the 77 countries are present in the data, so there is no selection bias at work in these patterns.

The differences between countries might be explained by institutional characteristics of the countries in which the ISPs operate. Two institutional differences that I tested for are: *i)* the development of the ICT infrastructure of each country and *ii)* the overall economic status of the country. In both cases we should expect to observe more victims in more developed countries, as more online activity and better infrastructure might drive more motives and opportunities for attacks – around online gaming, for example.

The ICT development index is a well established indicator of ICT development ranging from 1 to 10 with higher values for more developed countries which I have tested my data against. The index is provided by the ITU (the United Nations International Telecommunications Union) and constructed from a set of internationally agreed-upon indicators. I also looked at the gross domestic product at purchasing power parity (GDP PPP) per capita, to capture the economic status of each country [180]. From the plots in Figure 6.6a and Figure 6.6b, it is clear that both explanatory variables do correlate with ISP victim rates, but only weakly.

To consider the joint effect of the three explanatory factors that I have examined so far, i.e., the number of ISP subscribers, ICT and GDP PPP indexes, I construct several statistical models using negative binomial, generalized linear model (GLM) regression. The models predict the



(a) ICT development index (b) GDP PPP per capita

Figure 6.6: Correlation of ISP victim rates with country level development statistics, namely ICT development (left) and GDP PPP of the countries in which ISPs operate

number of victims per ISP given a set of explanatory variables. A summary of these statistical models are presented in Table 6.2.

$Model_1$ only includes the attack surface size, $Model_2$ adds the ICT development index as an additional factor and finally $Model_3$ adds the GDP PPP per capita. As expected, $Model_1$ demonstrates the effect of the size of the subscriber population – i.e., the size of the 'attack surface' – in correspondence with my earlier results (Figure 6.4). The other two models demonstrate that in addition to size, the two institutional country level variables considerably contribute to the variation in the number of victims per ISP, however their effects are much smaller. I interpret the results of $Model_2$ as an illustrative example. While holding everything else constant, increasing the number of subscribers by one unit (equivalent to multiplying the number of subscribers by 10 due to the log_{10} scale of the variable) multiplies the number of victims per ISP by $e^{1.996} = 7.36$. Similarly, increasing the ICT development index by one unit (while other factors are held constant) multiplies the number of victims by $e^{0.249} = 1.28$. $Model_3$ can be interpreted in a similar fashion. Note that due to the correlation of ICT development and GDP I did not include both variables in one model.

I have also examined other factors, such as 'gaming popularity' and 'piracy' which show weak correlations with victimization rates as well. Including these in separate GLM models shows a significant small effect of online gaming as captured by the average number of games owned per country on the Steam online gaming platform. This could be indicative of a possibly weak relation with online gaming and end-host victimization. However, further examination of the variable indicates strong correlations with ICT development and GDP therefore bearing little added information which the other factors did not already include in the models.

Table 6.2: Negative binomial GLM regression models with 'Log_e' link function for number of ISP victims

	Dependent variable:		
	# Victims per ISP		
Models →	(1)	(2)	(3)
↓ Independent vars.			
Subscribers	2.160^{***}	1.996^{***}	1.977^{***}
(log_{10})	(0.079)	(0.075)	(0.074)
ICT Dev. Index		0.249^{***}	
(2015)		(0.034)	
GDP PPP per Capita			0.030^{***}
(in $1000)			(0.004)
Constant	-5.880^{***}	-6.712^{***}	-5.705^{***}
	(0.454)	(0.468)	(0.430)
Observations	304	300	291
Log Likelihood	−2,255.880	−2,204.260	−2,128.202
θ	0.963^{***} (0.070)	1.097^{***} (0.082)	1.143^{***} (0.087)
Akaike Inf. Crit.	4,515.761	4,414.520	4,262.404
Note:			*p<0.1; **p<0.05; ***p<0.01

Given that the institutional factors have a weak effect, it begs the question of why, in the majority of the countries, ISP victim rates are closely clustered together. More specifically, the ISPs of only 12 of the 77 countries are dispersed by more than one order of magnitude (among them are Brazil, India, and China). Even with quite similar infrastructure and economic conditions, the differences among ISPs are larger between the countries than within them. This pattern suggests that there are specific country-level factors at work, beyond the general factors that I have examined.

We can only speculate why ISPs in a certain country are so clustered, but one explanation is that attackers and victims are geographically concentrated and that their interaction leads to network-effects. We know from the research on booters that many of the customers are gamers [169]. Other studies have told us that many of the victims are also related to gaming [149]. Combine this with findings from online social network analysis, inside and outside of gaming, which found that these online networks are shaped by geographical vicinity. In other words, users in online networks tend to be friends or familiar with each other in offline networks as well [181, 182]. In other words, they are geographically close.

Jointly, these three factors might drive a geographically concentrated network effect: some of the victims become attackers themselves, which

is easy because of the booter services. These new attackers, in turn, victimize others, and the cycle continues. This pattern fits with anecdotal evidence from news reports. In the Netherlands, for example, DDoS-ing became such a widespread phenomenon among schoolkids [175], that even those who did not play online games started to use booters, because everyone was doing it. One more technically skilled youngster said he quit DDoS-ing, as "it became too easy" and "even my sister can do it" [183].

Overall, these findings reveal that the number of subscribers of ISPs is a very strong predictor for the number of victims per ISP (see Figure Figure 6.4). This suggests that the chances of being victimized are surprisingly uniform across ISPs. The accessibility of DDoS-as-a-service might have caused a democratization of victims: everywhere there are now regular users deemed worthy of attack. This is a far cry from the highly publicized attacks on high profile targets like governments and enterprises. Those are attacked too, of course, but the bulk is targeted at regular netizens.

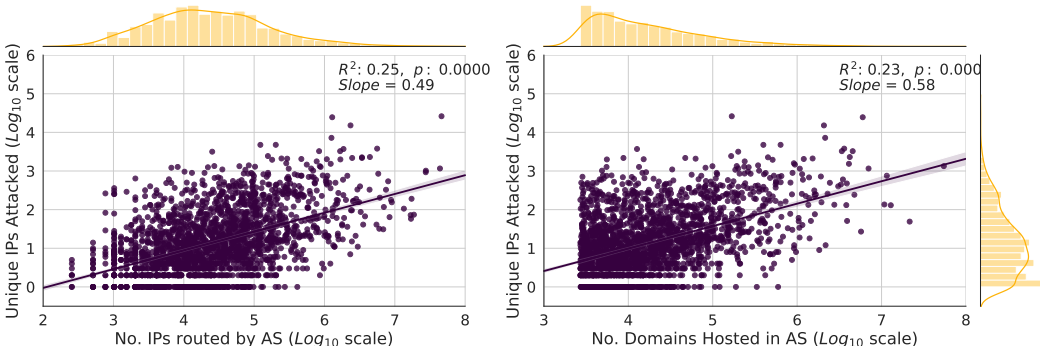
That being said, I do see significant variation in terms of victimization rates. The country-level differences are partially explained by institutional factors and partially by specific country-level effects. In the absence of direct evidence, I speculated that the remaining variation might be driven by geographically concentrated network effects.

6.6 HOSTING PROVIDERS

In this section I take a closer look at victims located in hosting provider networks. As for ISPs, the main questions at this stage are: *How are victims distributed across different hosting ASes* and *Do some hosting providers have disproportionately more victims than others?*. Unlike broadband victims, here, I do not expect the dynamic nature of IP allocation to significantly effect or lead to a misrepresentation of the number of victims as hosting networks do not typically exhibit IP churn. Therefore we can examine the distribution of victims over networks by simply counting the number of unique victim IPs that have been observe perAS.

As with broadband networks, I expect differences in customer base or network size to correlate with the number of victims. To test this, I need to estimate the size of the hosting providers. One approximation, which I have discussed in previous chapter as a measure of provider 'exposure', is to use the number of hosted 2^{nd}-level-domains (2LDs) per each provider. Recall however, that I found that only 46.6% of the hosting victim IPs have been observed to host domains. This implies that the number of domains will not be a very reliable approximation of the attack surface size in this instance. Therefore, we can also use the number of routed IP addresses by each hosting provider as a second proxy for its size to compare against. This metric, however, is less

able to account for shared hosting (several 2LDs sharing the same IP address). As I will see below, using both proxies in combination gives the best results.



(a) Routed IPs (b) Hosted domains

Figure 6.7: Correlation between hosting network victim counts and estimates of provider size based on number of routed IP addresses (left) and number of hosted 2LDs (right). Histograms along axes show distribution of data points along axis. Graphs share the same y-axis.

Figure 6.7a and Figure 6.7b plot the number of unique victim IPs per hosting provider against the number of routed IP addresses and hosted 2LDs of the provider respectively. Both figures demonstrate a moderate effect of attack surface size on the number of victims, but size does not appear to explain a large portion of the variance as indicated by the relatively lower R^2 values in comparison to what we observed in the previous section. This simply means that only a small part of the variation among hosting ASes is explainable purely through the attack surface size. But as before, we can see that some hosting ASes are disproportionately attacked more (data points far above the regression line) or less (data points far below the regression line) in relation to their size. This signals that attacks on hosting providers are also quite strongly driven by other explanatory factors. The question to consider then is *what additional factors can explain the variation that we observe after the size effect has been corrected for*? Again, as before correcting for size effects can be achieved through dividing the number of victims per provider by the size estimate of the provider.

One possible non-size related explanatory factor that I consider is related to the popularity of the hosted content. The expectation here is that more popular content is more likely to be attacked. In my analysis I use the list of top 1 million Alexa ranked domains as a proxy for the popularity of the hosted content [184]. Given the 2LDs that we have identified per hosting provider using pDNS data, I use the median ranking of the subset of top 1M Alexa ranked domains per

each provider as an indicator of the popularity of content that it hosts. Note that in my analysis I use reversed rankings: the most popular Alexa domain has the rank of 1,000,000 to make the interpretation of my results more intuitive.

A second possible factor that I consider is the type of hosting service that is offered. I expect that dedicated hosting is more likely to be attacked in comparison to shared hosting and other similar cheaper services offered by hosting providers. I use the number of IP addresses that have been used by the hosting provider to host all of its 2LDs as an indicator of the type of hosting. This indicator combined with size estimates (routed IPs and hosted 2LDs) captures the spread/density of domains per available IP address. A lower density of domains per IP is an indication for more dedicated services to their customers, while higher densities are indicators of shared hosting.

The analysis of these non size-related factors demonstrates a weak correlation with the number of victims per provider after correcting for size effects. For the sake of brevity I do not include the details and instead move on to consider and report on the joint effect of all explanatory factors through statistical modeling of the data instead.

In a similar fashion to what I did for broadband victims, I construct several statistical models of the number of victims per hosting provider using negative binomial GLM regression. A summary of these models is presented in Table 6.3. They clearly demonstrate that for larger attack surfaces there are more victims.

$Model_3$ uses all variables to explain the variance in victimization of hosting providers. Due to the unavoidable correlations between these variables I include interaction terms which control for the covariance between them. The model demonstrates that when considered jointly, the number of hosted 2LDs and the popularity of content have a significant effect on the number of victims per hosting provider. As expected, the size-related factor has the largest effect while the popularity of content as represented by the median Alexa rank is moderately affecting the victim numbers. It also suggests that there is not enough evidence to support the hypothesis that the density of domains or type of hosting has a significant effect on victim numbers. Due to the inclusion of interaction terms, $Model_3$'s results need to be interpreted in a slightly different manner. The more complex and improved model (as indicated by the improved log likelihood) suggests that while holding all other factors constant, increasing the 'Hosted Domains' variable by one unit (equivalent to multiplying the number of hosted 2LDs by 10 due to the log_{10} scale of the variable) multiplies the number of victims by $e^{1.050-0.338+0.198} = 2.48$. Increasing the 'Median Alexa Rank' variable by one unit (equivalent to multiplying the median Alexa rank of the content by 10 due to the logarithmic scale) multiplies the number of victims by $e^{0.305} = 1.35$. Finally, note that in $Model_3$ the number of routed IPs

Table 6.3: Negative Binomial GLM regression models with 'Log_e' link function for number of Hosting Victims

	Dependent variable:		
	# Victims per Hosting Provider		
Models →	(1)	(2)	(3)
↓ Independent vars.			
f_1: Routed IPs	1.198^{***}		0.507
(log_{10})	(0.040)		(0.354)
f_2: Hosted Domains		1.237^{***}	1.050^{***}
(log_{10})		(0.050)	(0.243)
f_3: IPs with Domains			-0.415
(log_{10})			(0.427)
f_4: Median Alexa Rank			0.305^{***}
(log_{10})			(0.075)
$f_1 \times f_2$			-0.338^{***}
(Interaction term)			(0.088)
$f_1 \times f_3$			0.266^{***}
(Interaction term)			(0.044)
$f_2 \times f_3$			0.198^{**}
(Interaction term)			(0.084)
Constant	-1.120^{***}	-0.988^{***}	-3.859^{***}
	(0.177)	(0.215)	(1.093)
Observations	2,203	2,203	2,203
Log Likelihood	−10,594.160	−10,703.310	−10,192.260
θ	0.421^{***} (0.011)	0.393^{***} (0.010)	0.546^{***} (0.014)
Akaike Inf. Crit.	21,192.330	21,410.620	20,400.520
Note:			*p<0.1; **p<0.05; ***p<0.01

is not a significantly contributing factor. This does not negate the size effect as observed in $Model_1$ and simply means that when considered jointly with the other factors the number of routed IPs does not add more information to the model that has not been already captured by the other included factors. Based on these results we can conclude that in addition to size factors, which have the strongest effect on the number of victims per hosting provider, the popularity of content also weakly contributes to this number.

To get a better sense of the actual victims, I have taken a closer look at some of the hosting victims that are associated with domain names. Many IP addresses are associated with multiple domains, obscuring the target and potential motive of the attackers. However, a subset of around 23,855 IP addresses are associated with just a single domain name according to our passive DNS data. I took a random sample of 1% of this set (238 domains) and checked all of them manually to assess what type of website was being attacked. Of the 238 domains, 107 no

longer showed any content. Most of them could no longer be resolved, others ran into connection issues or were replaced by parking pages. Given that the victim data was collected over two years, some degree of 'link rot' is to be expected, though this decay of domains is much higher than those found in other studies (e.g. [185]), suggesting that a lot of the victims had a somewhat fleeting presence on the web, rather than being well-established businesses or organizations.

Of the 132 sites that offered content, 55 sites (42%) were directly related to gaming. Of these, 27 were associated with a single game: Minecraft (17), followed by Counterstrike (6) and Runescape (4). The remaining 77 sites (58%) were highly heterogeneous, including but not limited to a few large stores, an airline, two football clubs, two schools, two escort services, one porn site, several hobby forums, a casino, a nature conservancy, and Twitpic, owned by Twitter since late 2014. In short: motives for DDoS attacks are highly varied, though gaming-related feuds are the most dominant of motives it appears. In the Minecraft community specifically, DDoS attacks seem to be part of the culture.

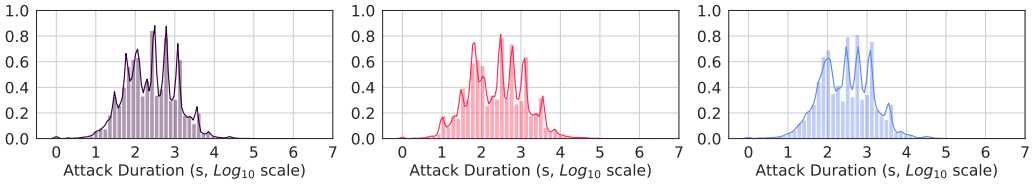
We can summarize the results with respect to hosting providers as follows. Hosting providers constitute the second largest group of victims in the amplification honeypot data. Some providers are attacked disproportionately more than others. This can be partially explained by the size of their attack surface. Furthermore, hosting popular content increases the number of victims. Finally, in agreement to what others have also found I see a large victimization of gaming related resources within the hosting provider networks.

6.7 ATTACK DURATION

In previous sections I have examined the question of who gets attacked more, whether that is disproportionate and if some factors can explain the variance among victim counts. We can also approach the question of who gets attacked more from the view point of time. That is, rather than looking at victim counts we can also approach the question as *who gets attacked longer and possibly why?*

To answer these questions, I take all victim IP addresses and measure the intervals under which they were continuously attacked. These intervals are calculated regardless of which AmpPot or protocol was used to attack the victim IP. The resulting interval lengths are defined as the attack duration. Note that here, I have merged attacks that are closer than 600 seconds apart and consider them as one continuous attack on the victim. Given these durations, the primary question is *whether the distribution of these durations differs per victim type*. These distributions are shown in Figure 6.8.

The median attack duration for broadband ISPs, hosting and the other types of victims are 272, 285 and 300 seconds, respectively. One surprising observation is the frequency of relatively short attack durations. The majority of attacks are shorter than 286 seconds long. For attacks longer than 300 seconds, I observe similar distributions of attack durations for all three types of victims. Interestingly, I observe an increased number of attacks that last around 5, 10, 20, 60, or 120 minutes which correspond to the peaks observable in the figure. The trend suggests that, in general, the attacks are largely originated from booter services and are most possibly driven by attackers' capabilities rather than victim types (see Figure 6.8).



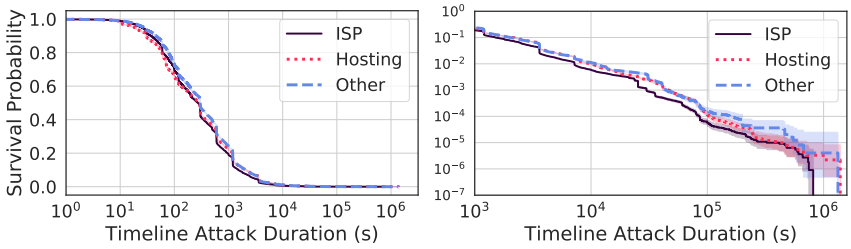
(a) Broadband ISP (b) Hosting (c) Other

Figure 6.8: Distribution of attack durations for various victim network types.

To further compare the differences in durations for different victim types, I use a well established statistical technique that is commonly referred to as survival analysis, which I also employed in **??**. The technique is used to answer questions about the proportion of a population that will survive past a certain point of time on a measurement timeline and at what rate the individuals 'survive' or 'die'. In our case, the event that we analyze is the '*end of an attack*' on a victim IP. Figure 6.9 demonstrates the survival analysis results. I use the Kaplan-Meier estimator to approximate the survival function [186], measuring the probability of an attack exceeding a certain duration for various victim types.

A log-rank comparison of the survival probabilities indicates a significant difference at a 0.99 confidence level between attack durations on different victim types. The log-rank chi-square statistic comparison between broadband/hosting, broadband/other and hosting/other are equal to 2,131.8, 3,493.4, and 739.3 respectively. These results indicate a significant difference among the attack durations per victim type, however in terms of magnitude, the differences seem to be quite small (see Figure 6.9).

We can also compare the survival rates of each victim type using the Cox proportional hazards model. The Cox model does not depend on distributional assumptions of survival time and allows to estimate the hazard ratio defined as the relative risk based on a comparison of event rates. The hazard ratios show that relative to hosting providers,



(a) All (b) 10% longest attacks

Figure 6.9: Survival analysis of attack durations

attacks end 14% faster for broadband victims while 3% slower for the other type of victims. While the results demonstrate that attacks are statistically shorter on broadband ISP victims the magnitudes of the differences are not large enough to have significant implications.

To conclude, all victim types experience attacks ranging from short lived attacks in the order of several seconds to long attacks which last several days. In other words, there is no significant variance among the duration of attacks on victims of different types.

I have further manually analyzed victim IP addresses of the 100 longest attacks of which 98 lasted more than 24 hours. They were launched against 87 unique IP addresses and 46 unique ASes. Interestingly, I do not observe any domains historically hosted on as many as 41 IP addresses (47 attacks). Of these, 6 IP addresses were directly related to gaming, including two victims against which the attacks lasted more than 16 days. Of the remaining 46 unique IP addresses, which were identified to be hosting some content, 17 were mapped to just a single domain name in passive DNS data. Of these, I have identified 6 victim IP addresses that hosted websites which provided torrent files to facilitate P2P file sharing, 4 websites related to gaming, 2 chat websites, one Internet banking website, and one TorGuard VPN website. By manual analysis of 15 IP addresses for which I observed 2 or 3 domains, I have further identified three victim IP addresses that mapped mainly to torrent, gaming, and TorGuard websites, respectively. The remaining 14 victim IP addresses mapped to more than 3 domains; 4 among them appeared to be used for shared web hosting and they mapped to 51, 346, 614, and 931 domains. To conclude, this manual analysis reveals that not only gaming but also torrent sharing-related IP addresses are among long-duration attacked victims.

6.8 RELATED WORK

Much research has been devoted to analyzing the technical properties of amplificationDDoS attacks: which protocols can me misused and how; how large the population of vulnerable reflectors is; how difficult or easy it is to find and misuse these reflectors; and how they could be mitigated [136, 139, 149, 187]. We know for example that many UDP based protocols are prone to be misused (`NTP`, `DNS`, `SNMP` and `Chargen`) and we know what their amplification factors are [139]. We also know how large the populations of vulnerable devices running these protocols are [139, 149, 168] and what kind of a threat they pose. Darknet and honeypot traffic reveals how perpetrators are actively scanning for such devices in the wild [139, 149, 166, 188]. Some have even attempted attacking their own infrastructure in order to asses the potential damage of booters and surprisingly find their damage to be much smaller than the spectacular cases reported in the news [175]. Others have examined the motives behind the provision of booter services through interviews [189]. Analysis of trends also reveals how over time, specific protocols rise and fall out of popularity among attackers and how remediation and intervention has affected the landscape [149, 171].

Earlier work on amplification DDoS attacks have focused less on studying the victims. The most in-depth understanding comes from the special case of `NTP` attacks, which allows probing the amplifier for victim IP addresses. Czyz et al. [149] provided the most comprehensive overview. The analysis of the smaller subset of victims from leaked booter databases [169, 170] also point towards gaming-related victims. In this chapter I corroborate earlier findings, especially [149, 171], that many of the victims are end hosts and gaming-related resources, but I also expand on this and show that the distributions have shifted. Moreover, I provide a wholly novel contribution by developing victimization rates and providing an explanatory analysis of key determinants behind victimization patterns.

Finally, part of what we know about victims is based on industry reports from DDoS mitigation providers [172, 173, 174, 176]. These mostly provide information on the type of industry that is affected most by DDoS attacks and point to the gaming industry and software industry as main victims. The results presented in this chapter paint a rather different picture, agreeing only with those reports in that many victims are gaming-related. Industry reports seems to be vulnerable to biases related to the fact that data collection often takes place in networks of the customers of DDoS mitigation providers.

6.9 DISCUSSION AND IMPLICATIONS

This study discussed in this chapter has presented the first in-depth look at victimization patterns of DDoS amplification attacks — and thus of the booter services that drive the bulk of these attacks. I found that broadband networks harbored most of the victims (62%), followed by hosting networks (26%). Educational, governmental and enterprise networks make up just a small fraction of the victim population (12%), contrary to industry reports and news items about high-profile attacks.

The population of victims is predictably distributed across broadband and hosting networks. To a large extent, the size of the user population drives the victimization rate – in broadband around 60% of the variance in victim counts can be explained from just the number of subscribers of the provider. Further explanatory factors are ICT development and GDP PPP per capita. I also see significant differences among countries however, that are not explained by these institutional factors. Remarkably, within most countries, ISP victimization rates are clustered together. This implies there are specific country-level effects at play, perhaps the result of geographically concentrated network effects among attackers and victims.

In hosting provider networks, the size effect is also visible, though less pronounced. The popularity of content, as measured by Alexa rankings, had a small effect. When I looked at victims IP addresses associated with a single domain, I found that 42% of the sites I could identify were related to gaming, most notably to Minecraft.

Attack duration did not differ significantly across the victim populations. When I examined the 100 longest attacks, 98 of which lasted more than 24 hours, I found, again, mostly gaming-related content rather than high-profile targets.

What do these findings mean for the consequences of the so-called commoditization of DDoS attacks? Rather than going after high-value targets, DDoS-as-s-service has invited attackers to go after regular users. With the commoditization of attacks, victimhood has democratized. And so has criminality, in all likelihood. Assuming that the users are targeted by someone that actually knows them, rather than by a random stranger, my findings imply that the attacker population has also broadened. In short, booters have indeed drawn more attackers into the DDoS ecosystem, as commoditization theory suggests, and this has led to a an expansion of victims among regular users, who now make up the bulk of all victims.

Overall, the fact that most victims are regular users suggests that profit is not a dominant motive anymore, assuming it ever was. The commoditization provided by booters has enabled attacks for as little as one U.S. dollar. This type of cybercrime is priced in the same range as, or even below, many entertainment products. It is now cost-effective

to pursue many more motives than profit, even very frivolous ones – like harassing your schoolmates during Minecraft games or online chats. Many of the new attackers probably do not see themselves as cybercriminals. Everyone is doing it, and they are not making any money from it.

The fact that attack patterns are so proportional to the number of users might seem unsurprising, but it has far-reaching implications. Rather than a phenomenon of motivated attackers with specific objectives and targets, DDoS has become a cultural phenomenon. The closest parallel to the observed pattern seems to be wide-spread use of torrents and file lockers to download copyright-infringing materials. This suggests a new route of action for fighting the DDoS problem: rather than using criminal law to go after motivated attackers, a better approach might be what criminologists call *situational crime prevention* [24]. It shifts the focus from identifying and penalizing attackers to taking away the opportunities that trigger crime. It can draw on a much broader mix of measures, often based on civil rather than criminal law. It can range from soft measures, such as awareness campaigns for youngsters, to harder ones, like the takedown of booter accounts by providers such as PayPal [171].

What are the implications of the findings for the wider commoditization of cybercrime? Should we expect an influx of attackers and an expansion of victims in other criminal markets as well? Not per se. As Florencio and Herley have argued, cybercrime is often harder than it looks and it scales less well than one would assume at first glance [87, 190]. Indeed, in many markets, we do not see the rapid expansion of crime that effective commoditization would cause. This can be explained by the fact that many of these service models do not supply complete criminal value chains. Take fraud using banking Trojans for example. It is one thing to buy malware-as-a-service and distribute it via pay-per-install, but that doesn't mean one can successfully execute online banking fraud. There are bottlenecks elsewhere, especially in the use of money mules and other cash-out channels. Mules-as-a-service did not manage to solve this bottleneck yet.

We see the predicted effects of commoditization in DDoS attacks, because here the booter provides the value chain end-to-end. In other forms of cybercrime this seems much harder or even impossible, though some might come close, like ransomware-as-a-service using bitcoin. And indeed, we did recently see an explosion of ransomware attacks. We can only hope that for many other forms of cybercrime, bottlenecks will remain resistant to successful commoditization.

Part III

CONCLUSIONS

7 CONCLUSIONS

The objective of my research has been to reduce the information asymmetry surrounding the security efforts of hosting providers. It seeks to measure and quantify how effective providers are in securing their infrastructure, in addressing the problem of harmful Internet content, and dealing with the negative externalities caused by the abuse of their services towards cybercrime. In other words create transparency around the security efforts of hosting providers. As such, the main question that it answers is as follows:

> *How can we quantify the effectiveness of hosting provider security practices?*

This research question was broken down into five smaller sub-questions which were then answered in each of the preceding chapters.

In this concluding chapter, I connect what I have learned in the previous chapters to my dissertation's objective and its main question, summarize the findings, and reflect on their implications for the hosting market, in addition to discuss broader implications for the more general problem of harmful Internet content and how the market may be governed towards more desirable outcomes. I conclude this work by highlighting some of its limitations, in addition to discuss directions that future research may take.

7.1 SUMMARY OF FINDINGS

At the onset of my studies, multiple industry and academic efforts to quantify and compare the security of hosting providers were identified. It became clear that these did not adequately address the challenges of empirically measuring the state of security in the hosting market as a whole. Of particular importance are questions of how to identify hosting providers, how to quantify their exposure to abuse, and how to characterize the various types of hosting services that they offer, which are all essential to draw meaningful comparisons of their security efforts. In colloquial terms to compare apples with apples.

Within my work I define two types of metrics for comparing the security efforts of hosting providers.

First are a set of metrics based on how frequently abuse incidents occur by counting the number of observed incidents for providers over a period of time. These metrics, which capture incident frequencies, are reflective of **proactive** security efforts since they signal how well

security incidents are prevented from occurring in the first place They capture the idea that if providers proactively secure their infrastructure, they should have to deal with less incidents, something which should be reflected in a lower number of abuse incidents.

The second type is a set of metrics based on how timely incidents are remediated by the providers over that period. These metrics, which capture incident remediation times, reflect the **reactive** security efforts of providers once incidents have already occurred. They capture the idea that once incidents occur, independent of how secure the providers' infrastructures are, more secure providers should remediate incidents faster.

The first three studies of my thesis mainly focus on how to develop metrics for each of these types of provider security effort. My next two studies then extend the work in two directions. I first examine where and how the application of metrics may fail by closely examining how Bullet-Proof Hosting (BPH) providers operate as a special case study. Next, I examine the broader impacts of hosting provider negligence by closely examining the phenomenon of Distributed Denial of Service (DDoS) for hire services, i.e. so-called "booter" websites, and study how broad the set of victims of such abusive hosted websites may be.

Overall, my thesis yields a number of technical and methodological contributions that improve the state of the art, in addition to arrive at several empirical findings about security in the global hosting market. I summarize these by drawing on a causal model of security incidents that was developed and discussed in Chapter 3. The model serves as a framework for understanding why security incidents, e. g., instances of hosted harmful content, concentrate around certain hosting providers. The framework allows meaningful inferences, and comparisons of the security efforts of providers to be drawn from the metrics that I develop.

To recap, the model (see Figure 7.1) stipulates that security incidents are caused by '*attacks*' that compromise hosting services in various ways. The extent to which such incidents may occur are moderated by two factors. First, a provider's '*exposure*' to attacks - also referred to as its *attack surface* - and second, the '*security efforts*' of a provider to combat attacks. The larger the attack surface, i.e. the more exposed the provider is, the more likely it is to suffer from attacks and thus incur losses from subsequent abuse incidents. One may think of exposure as the number of customers that are served, resources that are utilized or hosted, the types of hosting services offered to customers and even the pricing of hosting services to name a few examples. On the other hand, the more effective a provider's security practices and efforts are, the less likely it is to suffer security incidents. These may be thought of as procedures to vet customers, monitor content and infrastructure for signs of compromise or the placing of security and access controls on hosting infrastructure and other resources to name a few examples.

The causal relation between attacks, incidents, and the moderating effects of exposure and security effort are captured by the causal model reproduced in Figure 7.1.

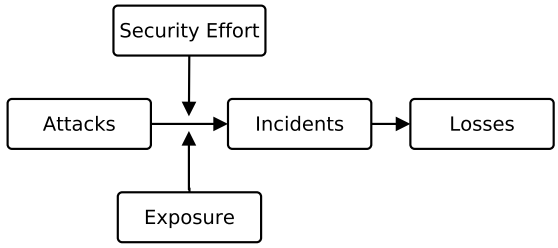


Figure 7.1: Causal model of security incidents for hosting providers, discussed in Chapter 3, which relates attacks to abuse events and the moderating effects of defender exposure and its security efforts to curtail potential attacks

Given this model, my research relies heavily on academic and industry efforts to collect security incident data that capture instances of abuse most notably blocklists. I use this data to extract new insights beyond the purpose for which the data is being collected. More specifically, I attribute discovered incidents to the specific providers that are responsible for hosting the discovered harmful content, and use it to empirically understand how attackers behave, understand the types and volumes of security incidents that providers experience, as well as infer how effectively hosting providers deal with abuse.

The first of my studies (Chapter 2) was concerned with developing metrics to compare the security efforts of hosting providers from empirical observations of abuse incidents. To this end, I employed data from seven so-called 'abuse feeds' and identified systematic steps to translate captured incident data into metrics reflective of hosting provider security postures. Based on the abuse data, I found abuse incidents to be mostly concentrated around a few hosting providers. Several stand-alone metrics, as well as an overall score, were then produced as point estimates, after taking into account the fact that more exposed providers are more prone to being abused. That is, once the exposure effects are explained away, this study assumed that differences among providers should be driven by how effective their security practices are and ranked providers on that basis.

Within the study I produced metrics to compare both the proactive and reactive security efforts of providers. Here, the analysis of the proactive and reactive security metrics suggested that these two dimensions of security effort are largely uncorrelated. This meant that only a small proportion of providers were observed to be both proactively and reactively effective in combating abuse (relative to their counterparts)

with most only performing relatively well on only one of these aspects. Two subsequent studies, discussed shortly hereafter, therefore each have focused on developing enhanced methods for producing metrics and empirically measuring each of the two dimensions of security effort independently.

In my first study, I adopted a pragmatic approach of identifying hosting providers from Autonomous System Numbers (ASNs). This was the typical approach taken within existing literature at the time. ASNs are technical identifiers in Border Gateway Protocol (BGP) Internet routing data signifying organizations that route Internet traffic to and from IP addresses. In other words each hosting provider is assumed to operate its infrastructure from within an Autonomous System (AS) with a designated ASN. I attributed abuse incidents to specific hosting providers by matching the IP addresses of servers on which harmful content was discovered against ASes that routed traffic to and from those servers. Moreover, by combining the same BGP routing data and the AS information which it contains with passive DNS (pDNS) data, and IP address geo-location data, I constructed a list of Dutch hosting providers, estimated their attack surface size (exposure), and characterized their hosting services via measurable quantities such as the number of IP addresses routed by, the subset of IPs associated with domains names, and the number of domain names operated from within each hosting provider's AS. One of the main contributions of this study is that it was a first-of-a-kind to develop techniques for measuring exposure based on pDNS data.

One unsolved problem in this study's approach was the fact that AS ownership information is not a perfect basis for identifying the provider responsible for hosting harmful content on a certain IP address. Within an AS, there can be multiple providers, each with their own IP space and server infrastructure. Ownership of IP address ranges (a.k.a. IP prefixes) is a better basis. And such IP prefix ownership is reflected in WHOIS data as it captures organizations to which different IP prefixes have been assigned to for use by various Regional Internet Registries (RIRs). WHOIS data also containts the abuse contact information for IP addresses, which also point to the providers who own the IP space, not to the AS owner who might only route Internet traffic to/from specific IPs.

Thus, a subsequent study, by my colleagues and myself [44], developed a more valid and reliable way of identify hosting providers at scale. The study resulted in improved methods for first, identifying hosting providers through publicly available WHOIS IP allocation data, and second, extended the use of pDNS data to derive additional information about the business models of hosting providers, for example, whether and how many of their hosted resources are on shared hosting services versus how many resources are hosted on dedicated infrastructure of their own. The former improvements enhanced our ability

[44] *Samaneh Tajalizadehkhoob et al. "Apples, oranges and hosting providers: Heterogeneity and security in the hosting market." In:* NOMS. *IEEE, 2016*

to identify a larger number of hosting providers most notably so-called hosting resellers - a significant yet smaller type of hosting business - that do not necessarily have a designated ASN. The latter improvements, allowed us to later also take into account how different types of hosting services, i.e. business models, influence the moderating effects of a provider's exposure to attacks.

One of the limitations of my original study was its implicit assumption that attackers may be seen as randomly attacking providers proportional to their exposure. Only by making this simplifying assumption about attacker behavior dynamics, would I have been able to attribute the differences in abuse concentration among providers to differences in their security efforts, of course after having taken exposure effects into account. That is, if attacker behavior is seen as randomly attacking providers proportional to their size, it should produce a heteroskedastic noise pattern in the observed incident counts which grows proportional to provider size. Thus by normalizing incident counts by provider exposure, heteroskedastic noise is transfomed into a random noise pattern equally affecting all providers, and therefore something which may be overlooked.

A second limitation was the fact that the metrics I produced were volatile point estimates of provider security postures as the metrics did not take into account the inherent noise of the abuse data itself. No error margins were produced to indicate the level of confidence that we may instill in the metric values. Thus, the metrics were subject to volatility, and even though sensitivity analysis suggested reasonable metric stability, this limitation was something that I improved upon in my subsequent studies.

My next study (discussed in Chapter 3), focused on developing enhanced metrics to compare the proactive dimension of hosting provider security effort. Similar to my previous study I employed several abuse feeds to collect incident data. Unlike the previous study however, the improved technique for identifying hosting providers from `WHOIS` data was used, thereby allowing to enlarge the scope of the study to a global setting as well as better identify hosting providers, most notably resellers. As such, irrespective of their type, more than 30K hosting providers were identified in the global hosting market which were analyzed in this second study. As before, pDNS data was used to quantify the exposure of the identified providers, in addition to identify the type of their business and services that they offered. The causal model of incidents that I discussed earlier was also more formally developed here. The model was used in combination with a Bayesian statistical inference technique and the collected incident data to derive a latent quantity reflective of the proactive security efforts of providers.

In this study I also took an additional step to examine the distribution of incidents over the global hosting market and found that attacks

may indeed be seen as randomly distributed over the global hosting market. While the brunt of the attacks were observed to be targeted at providers with larger exposure, the number of incidents for each provider were consistent with a random attack pattern and largely explainable through the exposure of the providers alone. This confirmed that the assumption about attacker behavior dynamics that I made in my previous study were reasonable.

The main contribution of this study was a set of enhanced metrics represented as statistical distributions (as opposed to the point estimates produced in the previous chapter) in addition to methodologically accounting for inherent biases and measurement noise that may be present in incident data. Figure 7.2 which reproduces the resulting metrics and comparisons of hosting provider security postures shows some of the results. I briefly summarize these results here again. The figure represents Individual hosting providers on the x-axis and their estimated security performance levels on the y-axis as distributions. Positive values on the y-axis reflect better security performance and negative values poorer performance compared to the other providers, while controlling for differences in exposure to the attacks. The mean value for each provider's estimated security performance distribution is represented by the thick black line. Colored regions around this line illustrate the $[2.5 - 97.5]$% credible interval of the distributions. The graph also distinguishes between providers for which incidents have been empirically observed or not, by representing their security performance distributions in orange and gray colored bands respectively.

As such, these metrics not only quantify the effectiveness of provider security efforts but also reflect measurement uncertainty, or in other words the level of confidence that we may instill in them. An important result of this study is that it I showed that the produced metrics are predictive of the number of incidents that providers experience over a fixed period of time. That is, I was able to explain between 78 and 99 % of the variations in provider incidents numbers for empirical data that was left out of the modeling phase to produce the metrics. Given these metrics, I found that a large number of hosting providers exhibit below average proactive security performance relative to their global counterparts, even when exposure effects such as size and the business models of hosting providers are taken into account. An interesting observation was that the produced metrics appeared to be better at reflecting provider failures as opposed to the opposite outcome of successfully preventing abuse. This is observable by the fact that the metrics exhibit higher confidence when providers performed poorly, i.e the credible intervals for the orange colored regions are smaller than the grey regions of the graph. This phenomenon is explainable by the fact that my models were unable to distinguish between three scenarios. First, that lack of incidents may be driven by good security practices,

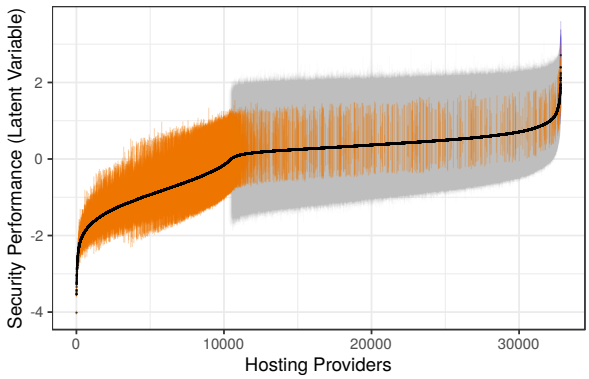


Figure 7.2: Comparison of estimated hosting provider security performance discussed in Chapter 3. Security performance is reported on a latent variable scale along the y-axis with positive numbers representing better security performance and negative numbers vice versa. All hosting providers identified across the market, are represented along the x-axis in improving performance order. 95% credible interval bands for security performance estimates are reported as color bands with mean values as solid black line. Providers for whom abuse has been empirically observed are orange color-coded and providers with no empirically observed abuse event represented in gray.

or second, driven by incomplete abuse data, or third, by having had small exposure. Nevertheless, the produced metrics, correctly reflect this uncertainty through wider spread credible intervals for the metric distributions.

Next, the study discussed in Chapter 4, focused on developing metrics to compare the reactive dimension of hosting provider security effort. I approached this by collecting incident data and comparing the amount of time required to take action against discovered harmful content on hosting provider networks. The approach yielded a more direct measurement of security effort and had the added benefit of its results not being affected by exposure effects. This meant that the developed metrics did not have to account for provider exposure to draw meaningful comparisons. Here, I find that complexities arise from remediation times being additionally influenced by actors other than the providers themselves, namely webmasters whose resources may have been compromised in addition to the attackers who may attempt to host harmful content directly themselves. As such, differences among how webmasters may react, and how attackers behave, have to be accounted for instead, to meaningfully compare the reactive security efforts of providers.

I first developed a framework to capture the various factors that influence remediation times based on the literature. Figure 7.3 which reproduces this framework summarizes the factors that may have an influence on remediation times. For example, within this framework I argue that if harmful content is directly hosted by an attacker it should be expected that he/she will not willfully cooperate in its removal thus prolonging remediation times. Moreover, certain types of content, for example Command-and-Control (C&C) centers, are typically more valuable assets which attackers will try to more carefully protect, more so than for example transient phishing pages and thus increase the security efforts required to take them down, i.e. increase remediation times. On the other hand, successfully taking down harmful content which has resulted from compromising resources assigned to legitimate hosting customers, depends on cooperation between providers and those customers and how well their security efforts combine. As such, the causal factors that influence remediation time are entangled.

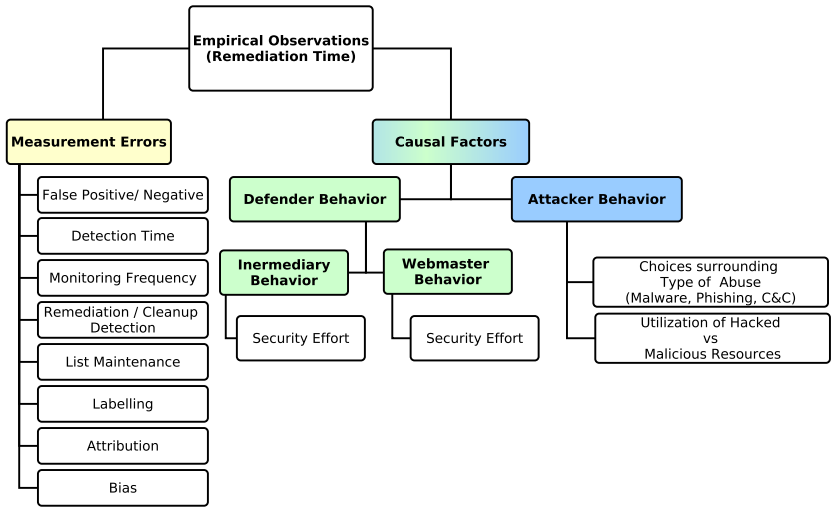


Figure 7.3: Analytical model of factors influencing abuse incident remediation times.

Given this analytical model, my study draws on statistical survival analysis techniques to model and explain the amount of time elapsed for individual incidents to be resolved while controlling for the various influencing factors to the extent possible. The statistical models which I employed controlled for differences in remediation time that may be attributed to differences between how webmasters and attackers may have behaved. These models allow one to compare and draw inferences about populations of providers and to infer which market segments have more vigilantly reacted to abuse and how their reactive security efforts relate to their structural properties such as size and

their business models. Here, I found that larger providers generally remediated incidents in a more timely fashion but that those with more shared hosting resources on average took longer to remediate abuse.

An second contribution of this study is that it demonstrates how widely used industry-sourced abuse data may be affected by various types of measurement error and how such errors threaten the validity of inferences that are drawn therefrom. I identified several such factors within my causal framework. These may be observed in Figure 7.3 under the heading measurement errors. Given their potential misleading effects, the study developed a methodology to triangulate results across multiple data sources and thus lessen the impact of measurement errors. As such I arrived at more robust results and could more carefully identify which providers had done a better job, in addition to provide explanations of why it was so. Alas, a number of findings in this study were found to be inconsistent after triangulation. This of course is to be expected, given the noisy nature of the incident data that was collected for this study.

My next study (Chapter 5) focuses on so-called Bullet-Proof Hosting (BPH) providers, a consistently difficult area of the hosting market to tackle. These knowingly allow miscreants to host harmful content and assist in its online persistence thereby enabling a large range of cybercrime. BPH providers are interesting cases to examine as they demonstrate the limitations of security metrics. They represents corner cases for which security metrics begin to fail, and exemplify how metrics, including ones that I have discussed in previous chapters, may be gamed and distorted. Understanding how such malicious providers operate is an important challenge due to their pivotal role in enabling cybercrime and may lead to better techniques for detecting and disrupting their operations. Therefore chapter Chapter 5, a first-of-a-kind study, undertakes a case-study of a recently taken-down BPH provider called `MaxiDed` by examining ground-truth data that resulted from the confiscation of its assets by law enforcement. The study provides a wide range of empirical insights into the inner workings of BPH providers.

While many BPH providers were initially detected and disrupted by relying on security metrics that reflect abuse concentration, I found that their operators have adopted agile operational models which render existing defenses and security metrics less effective. BPH operators heavily rely on reselling the hosting services of larger legitimate parenting hosting providers. They no longer necessarily own network assets with associated ASNs nor physical server infrastructure. At the same time, technical data that may potentially point to their organizations, for example WHOIS data, is often not up to date and thus not accurately indicate that certain IP prefixes may have been sub-delegated. That is, WHOIS data they still may reflect the original owners of the IP prefixes as the point of contact. Such inaccuracies lead to the concentrations of

harmful content hosted by BPH providers, to appear as originating from their often legitimate parenting hosting providers which sub-delegated their IP prefixes. Effectively, abuse emanating from their BPH services becomes diluted with that of their parenting organizations. I also found that BPH operators react to abuse complaints just as legitimate hosting providers would. That is, they forward abuse complaints to their customers and suspend servers that host harmful content if and when they are discovered, thus transferring risks of taking down abusive content to their potentially criminal customers. For such BPH providers, metrics that measure abuse concentration would often count abuse towards parenting hosting providers, while metrics that measure remediation times would typically observe reaction times that do not standout as outliers. Such are the edge cases for which security metrics fail to produce clear insights, demonstrating their limitations and how they are actively gamed and distorted by BPH providers in the wild.

A unique aspect of my study from Chapter 5 is that it also provided insights into the economics and profitability of BPH operations. I find that the profitability of BPH may be much more limited than what is widely assumed, but that other types of profitable cybercrime which it enables, may in conjunction, turn them into profit centers for their operators. The underwhelming profitability of BPH operations however, suggests that there may be other ways to tackle their menacing side effects, by targeting and applying economic pressure on their profits for example. For instance, through forcing them to operate under higher op-sec requirements thus increasing their operational costs and hopefully rendering BPH even less profitable. This of course remains to be seen.

My final study, discussed in Chapter 6, employs some of the methodological contributions of my previous work towards examining how security failures in the hosting market negatively affect society as a whole. It examines the negative externalities of so-called 'booter' websites. Booter websites, package and sell the ability to kick arbitrary targets offline by launching Distributed Denial of Service (DDoS) attacks at the click of a button. In essence they provide DDoS-as-a-Service at often low prices to anyone that is willing to pay regardless of their technical abilities. The empirical insights gained in this study demonstrate the extent of harm that malicious content may create if not adequately addressed, particularly by hosting providers that host them and are in a unique position to take them offline.

This study revealed that the bulk of the targeted victims of DDoS attacks emenating from booter websites are regular Internet users, most notably online gamers. I find that victimization rates across broadband Internet Service Provider (ISP) networks, i.e. the intermediaries that provide Internet connectivity to the victims, is highly proportional to the number of their subscribers. In fact, 60% of the variation in

victimization rates across broadband ISPs is purely explainable by the number of their subscribers. The observed pattern demonstrates how often such attacks take place and target regular Internet users. I also find that broadband ISPs from certain countries relatively experience more attacks that others. This is only weakly related to those countries having better developed ICT infrastructure.

Examining other types of vicitims in this study also revealed that hosting providers themselves, and their content hosting servers, are also occasionally attacked, although to a much lesser extent. Thus in terms of negative externalities, the providers hosting the booter infrastructure do not suffer the full cost of the DDoS attacks. In fact, most of the impact is suffered outside the hosting sector itself. This clearly demonstrates that the negative externalities that are caused by lax security practices of providers. Here I found that attack rates are proportional to provider exposure, for example the number of domains that they host or number of servers that they operate. A weak relation between the popularity of the hosted domains and suffering more attacks was also observed.

In terms of how damaging the DDoS attacks may have been I compared attack duration across both broadband ISP networks and hosting providers. Here I found both types of victims to experience similar attack duration. Longer attacks of course result in longer service disruption and are more costly to defend against by for example, activating protective measures such as traffic scrubbing.

In summary, the overall contribution of the set of studies is to improve our understanding of security in the hosting market. They have helped identify market areas where current security efforts fail, in addition to demonstrate the impact of such failures. Through answering the question of how the security performance of hosting providers may be measured and compared, we are provided with a set of empirical metrics as tools with which to identify the problematic areas and track progress. In what follows, I expand on the implications of this work and how security metrics may be employed to govern and steer the hosting market towards more desirable outcomes.

7.2 IMPLICATIONS FOR GOVERNANCE

In introducing my work, I have discussed how existing regulation has shaped hosting provider security behaviors by establishing a legal baseline of required security effort. I have also discussed how provider economic incentive structures have lead the hosting market to fail in addressing the problem of harmful content. Current regulation does not place liabilities on hosting providers when their services are abused to host harmful content. As a result, providers have mostly shifted burdens on to their customers or other 3rd parties to deal with abuse. While some providers take proactive and reactive countermeasures

against abuse, many, as I have demonstrated via the security metrics that I develop, do not adequately address the problem and exhibit insufficient security effort relative to their counterparts, even after taking other abuse influencing factors such as exposure and attacker behavior into account.

Currently, combating harmful content heavily relies on voluntary notice and take-down regimes through which third parties notify hosting providers of abuse and providers are in turn expected to take action against [53, 54]. In turn, providers typically notify the customers that are responsible from their point of view. Providers mostly forward abuse complaints with the expectation that customers would then take down the harmful content. While in my own studies I have found limited anecdotal evidence of some providers playing a more active role, few providers appear to strictly enforce security countermeasures or implement recommended anti-abuse policies to dissuade repeated abuse of their services. The problem is exacerbated by the fact that abusive customers may also migrate to other hosting providers with lax security practices. Ultimately, if providers or their customers do not take action against harmful content, court orders and law enforcement agencies may compel them to take action as a last resort, albeit a costly and non-scaling solution.

In short, the abuse of hosting services constitutes a collective action problem which involves multiple actors, with a key actor, namely the hosting providers themselves, not being adequately incentivized to engage with it.

In this section, I reflect on how the security metrics produced in this work, may support governance approaches seeking to incentivize provides towards taking more effective steps against abuse. These approaches may be categorized into four schemes typically identified in governance literature [191, 192, 193, 194]: (**i**) community-driven efforts, (**ii**) network-based efforts, (**iii**) self-regulatory or market-based approaches and finally, (**iv**) hierarchical governance approaches. I will reflect on each approach separately indicating how security metrics may contribute within each governance scheme.

We have already seen the emergence of *community-driven* efforts from within the hosting provider industry sector to address the negative externalities of harmful content. The Messaging Malware Mobile Anti-Abuse Working Group (M3AAWG) and its proposed best security practices for hosting providers which I discussed in Chapter 1 is a clear example of such efforts. The M3AAWG community, with a clearly stated aim of combating abuse, supports hosting providers by giving a wide range of recommendations on how to prevent abuse, how to detect and identify when/where abuse takes place, in addition to suggest ways to remediate it. Hosting providers may voluntarily follow a subset or all of these recommendations. Within such community driven

efforts, security metrics may help identify the security practices that are most effective by comparing providers based on the recommendations that they follow, whether it be through natural experiments, quasi-experiments or randomized control trials. As such, metrics may provide empirical grounding for prioritizing security recommendations based on how effective they are along with cost benefit and risk analysis. Security metrics also allow providers to individually track their progress when new security practices are taken on-board, for example by comparing to their own previous metric standings or in comparison to other comparable providers that have not implemented similar countermeasures.

We have also seen the emergence of *network-based* governance efforts to combat abuse. For example in the Netherlands initiatives like the `abuseplatform.nl` have brought together a network of organizations with trust, reputation, or business relationships in an effort to collectively commit to combating abuse. In this instance for example actors across government, academia, industry (i.e. several ISPs and hosting providers), in addition to several other stakeholders have come together and committed to sharing abuse data and tracking security progress. Members are asked to adopt a code of conduct that is based on the M3AAWG best practices document. Joining the platform also means that providers have to adopt this code. This particular platform for instance, has directly employed the security metrics that have been developed in this work. Here, the metrics are being used to privately communicate each participating organization's security stance relative to their counterparts. As such they raise self-awareness, and function as a means to incentivize each organization towards improving their security efforts. Other local or non-local network based governance initiatives may take advantage of existing network relations and employ security metrics similarly to raise awareness, track progress and incentivize hosting providers to implement more effective countermeasures.

Among the various governance approaches that I have mentioned, *self-regulatory* market-based ones have probably been the most tried out scheme as of yet. Security metrics, such as the ones produced in this work, have often been at the center of these efforts. Some metrics have been used to publicly name and shame bad hosting providers thus affecting their market reputation in teh hope that reputational effects may compel bad hosting providers to act against abuse. Other efforts are being tied to cyberinsurance schemes as alternative means of incentivizing market players to combat abuse. The premise being that a poor security posture may justify setting higher insurance premiums. In essence, security metrics are being used here to communicate that certain hosting providers are to be avoided, or for example to signal to customers, potential business partners, investors, and shareholders that some providers' services are excessively abused by cybercriminals.

Within this governance scheme naming and shaming of bad hosting providers have had some demonstrable effect in disrupting BPH operations for example. The use of security metrics to name and shame `McColo Corp.` and the `Russian Business Network`, both instances of BPH operations, is a concrete example which resulted in these provider's business partners severing their business ties. The use of security metrics within other efforts, for example `hostexploit` or Google Safe Browsing's transparency report, which publicly list hosting providers with high concentration of abuse are additional examples of naming and shaming efforts. In general however, the effectiveness of employing security metrics to communicate market signals are not yet clearly understood. Moreover, there has been certain pushback from the hosting industry based on the fact that security metrics do not take provider exposure into account, or based on arguments that the underlying methodologies and data are not transparent. By taking provider exposure into account, and openly communicating our methodology and underlying data of this work, some of these concerns may be alleviated. In fact, in communicating our metrics with Dutch hosting providers we have observed a more accepting tone. With respect to alternative approaches that go beyond naming and shaming, for example cyberinsurance schemes, I am not currently aware of any concrete examples beyond theoretical discussions of this approach. As such the effects of using metrics to determine insurance premiums still remains to be seen.

Finally, we have the category of *hierarchical* governance approaches which rely on existing authorities to further regulate the hosting market. Currently ongoing discussions of placing liability on Internet intermediaries within the EU, which I have pointed to earlier in Chapter 1, are clear examples of such hierarchical governance approaches. Within this context, security metrics may provide empirical basis to, and help policy makers in proposing evidence based policies. A second concrete example is the case of the `CleanNL` initiative in which Dutch law enforcement directly employed our metrics to identify the top 10 worst Dutch hosting providers and engage in direct talks with them. These efforts were a precursor to and may have had a positive effect on several of these providers joining the `abuseplatform.nl` initiative that I discussed earlier.

In short, security metrics may be employed to complement a wide range of governance approaches to tackle the collective action problem of hosting service abuse. While my own studies have largely focused on developing the metrics themselves, the extent to which they may complement each approach still remains to be seen. It goes without saying of course, that security metrics also have their limitations. These limitations and potential future research directions are topics which is will discuss next.

7.3 LIMITATIONS AND FUTURE WORK

Each of my studies discuss limitations and potential improvements that future research may undertake within the context of the individual study. Overall, these limitations stem from shortcomings in the empirical data that I used, methodological choices and assumptions that I made, as well as certain theoretical limitations, in addition to ethical considerations.

Here, I will briefly discuss such limitations and place them within the broader context of my research. I also discuss how future work may advance this work, in addition to how ongoing research in areas closely related to the problem of harmful Internet content may benefit from some of its results.

Limitations in Data

My research relies on a range of third-party data sets of which I had a black-box system view. These include a variety of abuse feeds that capture security incidents, Internet operations data, statistical data about companies, in addition to data collected from honeypot systems. An opaque view of their collection methodologies inevitably lead to limitations that are invariably linked to questions of data quality such as measurement error, coverage and bias.

A clear example of limitations that I encountered in this respect were challenges in utilizing Internet operations data such as WHOIS. A well known problems with such data is that it may be incomplete. For instance, for privacy reasons, portions of public WHOIS data are withheld, a fact that is also misused by miscreants [72, 74, 75]. Additionally WHOIS IP prefix allocation data may be inaccurate particularly with respect to IP prefix delegations [37, 44, 100]. Such inaccuracies not only affect the attribution of security incidents to specific hosting providers but also limit our ability to identify and enumerate the hosting providers themselves, for instance those that resell hosting services which typically operate from delegated IP prefix ranges. Notwithstanding its limitations, the use of WHOIS data constitutes a methodological improvement to the state-of-the-art in detecting and measuring security within the hosting market.

Data coverage issues were also encountered in for example using DNSDB's passive DNS (pDNS) data and the numerous abuse feeds that I employed in my studies. For the pDNS data, which I largely used to the estimate hosting provider exposure to attacks and to characterize their services, comparisons were drawn with a limited sets of authoritative domain registration data to further understand the coverage question. I found DNSDB data to provide reasonable coverage of the total number of domains registered within several popular generic top-level-domains,

for example .COM and .ORG in addition to several country top-level-domains e.g. .NL [112]. As such, I found certain guarantees that using DNSDB data results in only modest inaccuracies. With respect to data from abuse feeds, similar coverage issues were anticipated. Yet, beyond theoretically predicting how much of the existing abuse is covered by abuse feeds, there are no practical ways of further understanding coverage issues here as authoritative ground-truth data on the amount of global abuse does exist. Even industry leading efforts to monitor for abuse can only achieve very partial coverage at best and typically have little overlap with data collection efforts of others [109, 114]. Hence, my approach has been to employ data from multiple reputable and widely used abuse feeds in anticipation of potential coverage issues. The use of multiple higher quality and widely used abuse feeds in each study not only increase coverage, but also ensure that my findings have undergone corroboration and triangulation with existing literature that also analyze the same or similar data, in addition to provide safeguards against overtly confident claims based on low quality and low volume data. Moreover, as long as the coverage of these feeds does not bias against certain provider types quantifying security performance from this observational data should still be possible.

Yet in some cases I have encountered evidence of bias in a subset of the abuse feeds that I used. For instance, when studying incident remediation times, I found that irregular and infrequent monitoring of certain hosting providers, may have resulted in measurements errors that deviate from a random pattern thus indicating the existence of biases. More specifically, I found that my data sources may have monitored certain data subjects more frequently than others thus leading to biased measurements. In such cases I had to explicitly triangulate my findings across multiple data sets to alleviate concerns about the results being impacted by biased data. For some of the other abuse feeds that I used in other studies, I lacked reliable information on their collection methodology. However I also did not find clear evidence of bias within these data sets based on the analysis that I did. Yet, unknown biases may still exist which my analyses did not uncover.

Similar concerns of bias in honeypot data, which I used to study DDoS attack victims, existed. In this instance, I found the honeypot data to provide a less biased view of the DDoS attack phenomenon since my findings were more aligned with findings from several other academic studies that had also briefly touched upon the victimization question based on other data sets. In contrast, comparisons that I drew with DDoS protection industry reports on victimization patterns suggested that such reports were mostly based on data collected from a limited number of customer networks of the industry itself and as such resulted in a more biased understanding of the phenomenon. As such the use of honeypot data appeared to be a more sound theoretical choice.

Overall, I have encountered various limitations in the empirical data that I employed in my studies. While my approach of dealing with these has largely been one of methodological grappling with the data itself, for example removal of erroneous data points or triangulation across multiple data sets, there are limitations to how issues of measurement error, coverage and bias may be addressed methodologically. In light of these issues, future research that is based on third party data sets may emphasize more explicit corroboration and triangulation across studies, to ensure that findings are robust and not affected by the idiosyncrasies of particular data sets. To address data quality issues more directly however, future work may emphasize a more direct approach of dealing with the limitations. Collaborations with industry partners and third party data providers, should provide a much clearer understanding of the data generation processes behind each data set and help minimize limitations by improving data collection processes. As such the black-box perspective of the data which I had, may be opened up to produce a more transparent understanding of collection processes in addition to help address some of the limitations that I encountered more directly.

Methodological Limitations

In addition to the data-related limitations, there are also methodological as well as certain theoretical limitations to my work.

A prime example hereof, which is also linked to the limitations of WHOIS data that I discussed in the previous subsection, is the fact that even with perfect WHOIS data, identifying hosting providers is challenging. It is well known that due to the non-standard way by which WHOIS data is structured, parsing and utilizing it is hard [195]. This is why I have opted to use commercial WHOIS data from `MaxMind` which already deals with some of this complexity. There is also the problem that reselling of hosting services can take place without an agreement with an upstream provider, hence having no reflection in WHOIS data as I have seen in examining BPH in Chapter 5. Therefore there are methodological limitations to how much of the global hosting market may be identified and enumerated through such data.

A second example is an incomplete causal understanding of factors that drive the abuse of hosting services. The causal model of security incidents that my work relies on, is by no means a comprehensive one. There may be other causal factors that are currently unknown. At the same time, a range of causal and moderating factors exist that I have not been able to obtain data on because of practical limitations. For example, the pricing of various hosting services is theorized to be an influencing factor in provider exposure which I have not explicitly modeled or considered in my studies. That is, cheaper hosting services are more prone to abuse as attackers are rationally understood to minimize

their operational costs. Collecting data on the price of various hosting services was however infeasible at the scale of my studies. Similarly, collecting data on the specific security countermeasures that hosting providers may have adopted, a moderating factor in determining the effectiveness of provider security practices, was likewise infeasible from a scalability and external measurement perspective.

There are also certain limitations that stem from implicit or explicit assumptions that I have made in my work. For example, my initial study assumes that attacker behavior dynamics may be abstractly seen as a random noise element after provider exposure effects have been taken into account. And thus, due to their random noise characteristics, may be seen as naturally occurring random measurement noise which may be overlooked since statistical modeling techniques are well equipped to deal with such random error patterns. Lack of empirical data on attacker behavior is of course a practical limiting factor here. Also note that while my later studies empirically test the validity of this assumption, it is nevertheless a simplifying abstraction to enable the use of certain statistical data modeling techniques.

There are also limitations in the time frames of my studies. My studies are retrospective and observational in nature and by extension the metrics that they produce are too. That is, the produced security metrics do not reflect security performance in real-time but rather based on observations from time frames in the past. Therefore, to observe the effects of implementing new security countermeasures by particular hosting providers, new incident data needs to be collected and effects may only be observed with a time lag.

I should also note that certain fundamental theoretical limitations exist as well. In principle, 'security' is not something that can be directly measured in the absence of security incidents. In other words, what metrics measure is more describable as the lack of security which is indirectly observed through its manifestation as abuse and security incidents. This is a fundamental limitation which I have encountered for example when no incidents pertaining to certain hosting providers were empirically observed in the abuse feeds. A lack of security incidents here, of course do not imply perfect security, but may be attributed to a number of other reasons, for example small exposure or incompleteness of the incident data.

That being said, there are also limitations in the statistical techniques that I have employed to model incident data and explain variations in security incident frequencies and the reaction times of hosting providers. The use of Generlized Linear Models (GLMs), or Bayesian models in conjunction with Item Response Theory, and even survival analysis techniques may have had theoretical justifications, yet there is still significant variance among hosting provider responses that remain unexplained. This unexplained variance is linked to both limitations in

our theoretical understanding of the causal drivers of (in)security as well as a lack of empirical data.

Given such methodological limitations, future work may build a more comprehensive causal understanding of the factors that drive insecurity in the hosting market. Efforts to collect empirical data on additional causal factors may also help to better explain the variance among hosting providers and their responses to security incidents that I have observed. Specifically, future research may focus on developing more advanced statistical modeling techniques in addition to undertake longitudinal analysis to better understand additional factors that drive the abuse of hosting services and how security in the hosting market evolves over time as new security countermeasures are adopted by certain hosting providers.

Ethical and other Scientific Considerations

Several additional considerations regarding my research are also noteworthy. These relate to the type of data that I have partly employed.

For example, to study the behavior of Bullet-Proof Hosting (BPH) providers I have relied on sensitive forensic data that resulted from the confiscation of servers by law enforcement. This data contains personally identifiable information (PII). As such, its use may raise ethical concerns. It should be noted that In researching this data I have followed ethical guidelines set forth by the research community in the Menlo Report [157] as well as obtain permission from several entities, namely TU-Delft's ethical board (HREC), Dutch law enforcement and prosecutors. Note, however, that for additional ethical considerations I have limited my research to only portions of this sensitive data, and have avoided analyzing or reporting on parts of the data that contain PII. As such the analysis is only partial. Moreover, due to its sensitive nature, the underlying data is not openly available for others to replicate my results. To alleviate potential concerns in this respect I have carried out and reported on several extensive sanity checks.

I have also relied on several other froms of proprietary or commercial data sets which create reproducibility limitations. These data sets include DNSDB's pDNS data, WHOIS data from `Maxmind` as well as market analysis data on ISPs from `Telegeography`. The former two datasets are either available under academic license agreements or have openly available alternatives either through the same data provider or from other sources. Market data from `Telegeography` on the other hand is only commercially available.

Future Research Directions

My research has focused on the problem of harmful hosted content in relation to the abuse of online hosting services. The types of harmful content that I have examined are admittedly limited to ones that are more clearly distinguishable as such. For example, phishing pages, malware, botnet command and control backends, child sexual abuse material, and booter websites, to name a few. The metrics and methodologies that I have developed are, however, independent of the types of abuse and security incidents, and are capable of incorporating additional forms of abuse or vulnerability data to extend the research. As such, future work may consider a wider range of harmful content types particularly types which I have not considered here to produce more accurate metrics of comparison.

With that said, abuse and the problem of harmful content appear to be endemic to many other types of online services as well. For example, domain name registration services are also regularly abused by miscreants. Some of my other co-authored research has thus focused on developing security metrics for these other types of Internet intermediaries which do not fit within the focused context of this work [11, 112].

Such endemic abuse problems, combined with information asymmetry about the security efforts of a wide range of Internet intermediaries, have lead to wider spread calls to increase transparency around security efforts. For example, on the 30th anniversary of the Web, its inventor has called for governments to propose and enact regulation that is more suitable for the digital age, as the Web has also created opportunities for maligned groups to spread other forms of harmful content, for instance disinformation and hate speech which are affecting our societies [50]. Targeted political advertising on walled-off social media platforms have started to effect election outcomes, and extremist ideologies are being openly advertised in our online discourse. While determining which of the large volumes of online communication content are harmful is not as clear cut as the types of harmful content that I have considered, nonetheless the need to reduce information asymmetry about the security efforts of other Internet intermediaries is becoming also more evident.

Against the back drop of recent news, numerous governments are starting to demand other Internet intermediaries to also take more effective steps in combating harmful online content whether they be more clearly harmful examples such as phishing pages or less clear content such as hate speech. Within the EU for instance, there are deals that require social media platforms to remove hate speech by strict deadlines [196, 197, 198]. As such, there are clear connections between the remediation time metrics proposed in this work and what regula-

tors are calling for in such instances. To hold social media platforms accountable, reduce the existing information asymmetry about their security efforts, and increase transparency, future work in this area may be able to extend some of the methodologies developed in this work to measure and quantify efforts in these new problematic areas as well. However, clearly and justifiably so, civil rights advocates have also expressed concerns about having Intermediaries policing online content and the effects it may have on freedom of expression [199] as determining certain content as hate speech is not as straight forward as say identifying a botnet C&C. Therefore, in these instances a more nuanced approach is certainly required. Empirical metrics may help guide this discussion.

In conclusion, my hope is that the metrics and methodologies that are developed in this work may help reduce the negative externalities of cybercrime. I hope that through their application, hosting providers will be incentivized to more effectively combat the abuse of their services.

BIBLIOGRAPHY

[1] Karine Perset. *The Economic and Social Role of Internet Intermediaries*. Tech. rep. April. OECD, 2010. DOI: 10.1787/5KMH79ZZS8VB-EN. URL: http://www.oecd.org/dataoecd/49/4/44949023.pdf.

[2] Nick Nikiforakis, Wouter Joosen, and Martin Johns. "Abusing locality in shared web hosting." In: *EUROSEC*. 2011. DOI: 10.1145/1972551.1972553.

[3] Samaneh Tajalizadehkhoob, Carlos Gañán, Arman Noroozian, and Michel van Eeten. "The Role of Hosting Providers in Fighting Command and Control Infrastructure of Financial Malware." In: *ASIACCS*. 2017. DOI: 10.1145/3052973.3053023.

[4] Mark Felegyhazi, Christian Kreibich, and Vern Paxson. "On the potential of proactive domain blacklisting." In: *USENIX LEET*. 2010.

[5] H Liu, K Levchenko, M Félegyházi, Christian Kreibich, Gregor Maier, Geoffrey M. Voelker, and Stefan Savage. "On the effects of registrar level intervention." In: *USENIX LEET*. 2011.

[6] Shuang Hao, Georgia Tech, Nick Feamster, and Georgia Tech. "Monitoring the Initial DNS Behavior of Malicious Domains." In: *IMC*. 2011, pp. 269–278.

[7] Shuang Hao, Matthew Thomas, Vern Paxson, Nick Feamster, Christian Kreibich, Chris Grier, and Scott Hollenbeck. "Understanding the domain registration behavior of spammers." In: *IMC*. 2013, pp. 63–76. DOI: 10.1145/2504730.2504753.

[8] Janos Szurdi, Balazs Kocso, Gabor Cseh, M Felegyhazi, and C Kanich. "The Long Taile of Typosquatting Domain Names." In: *USENIX Security* (2014).

[9] Leyla Bilge, Engin Kirda, Christopher Kruegel, and Marco Balduzzi. "EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis." In: *NDSS*. 2011.

[10] Davide Canali, Marco Cova, Giovanni Vigna, and Christopher Kruegel. "Prophiler : A Fast Filter for the Large-Scale Detection of Malicious Web Pages." In: *WWW*. 2011, pp. 197–206. DOI: 10.1145/1963405.1963436.

[11] Maciej Korczyński, Maarten Wullink, Samaneh Tajalizadehkhoob, Giovane C.M. Moura, Arman Noroozian, Drew Bagley, and Cristian Hesselman. "Cybercrime after the sunrise: A statistical analysis of DNS abuse in new gTLDs." In: *ASIACCS*. 2018. DOI: 10.1145/3196494.3196548.

[12] Brett Stone-gross, Thorsten Holz, Gianluca Stringhini, and Giovanni Vigna. "The Underground Economy of Spam: A Botmaster's Perspective of Coordinating Large-Scale Spam Campaigns." In: *USENIX LEET*. 2011, pp. 4–4. DOI: 10.1371/journal.pone.0016637.

[13] Kirill Levchenko, Andreas Pitsillidis, Neha Chachra, Brandon Enright, Mark Felegyhazi, Chris Grier, Tristan Halvorson, Chris Kanich, Christian Kreibich, He Liu, Damon McCoy, Nicholas Weaver, Vern Paxson, Geoffrey M. Voelker, and Stefan Savage. "Click Trajectories: End-to-End Analysis of the Spam Value Chain." In: *S&P*. IEEE, 2011, pp. 431–446. DOI: 10.1109/SP.2011.24.

[14] Rob Thomas and Jerry Martin. "The underground economy: priceless." In: *USENIX; login* (2006).

[15] Jason Franklin, Vern Paxson, Adrian Perrig, and Stefan Savage. "An inquiry into the nature and causes of the wealth of internet miscreants." In: *ACM CCS*. 2007, pp. 375–388. DOI: 10.1145/1315245.1315292.

[16] Elias Raftopoulos and Xenofontas Dimitropoulos. "Detecting, validating and characterizing computer infections in the wild." In: *IMC*. 2011. DOI: 10.1145/2068816.2068820.

[17] *Google Safe Browsing*. URL: https://safebrowsing.google.com/.

[18] *Google Safe Browsing Transparency Reports*. URL: https://transparencyreport.google.com/safe-browsing/overview.

[19] Marie Vasek and Tyler Moore. "Do Malware Reports Expedite Cleanup? An Experimental Study." In: *USENIX CSET*. 2012.

[20] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. "A Nearly Four-Year Longitudinal Study of Search-Engine Poisoning Categories and Subject Descriptors." In: *ACM CCS*. 2014, pp. 930–941. DOI: 10.1145/2660267.2660332.

[21] Orçun Çetin, Mohammad Hanif Jhaveri, Carlos Gañán, Michel van Eeten, and Tyler Moore. "Understanding the role of sender reputation in abuse reporting and cleanup." In: *Journal of Cybersecurity* 2.1 (2016). DOI: 10.1093/cybsec/tyw005.

[22] Elie Bursztein, Borbala Benko, Daniel Margolis, Tadek Pietraszek, Andy Archer, Allan Aquino, Andreas Pitsillidis, and Stefan Savage. "Handcrafted Fraud and Extortion." In: *IMC*. 2014, pp. 347–358. DOI: 10.1145/2663716.2663749.

[23] Tyler Moore and Richard Clayton. "Evil Searching : Compromise and Recompromise of Internet Hosts for Phishing." In: *Financial Cryptography and Data Security*. 2009, pp. 256–272.

[24] Kurt Thomas, Danny Yuxing, Huang David, Thomas J Holt, Christopher Kruegel, Damon Mccoy, Elie Bursztein, Chris Grier, Stefan Savage, and Giovanni Vigna. "Framing Dependencies Introduced by Underground Commoditization." In: *WEIS*. 2015.

[25] Kurt Thomas, Rony Amira, Adi Ben-yoash, Ori Folger, Amir Hardon, Ari Berger, Elie Bursztein, and Michael Bailey. "The Abuse Sharing Economy: Understanding the Limits of Threat Exchanges." In: *RAID*. 2016.

[26] Brett Stone-Gross, Marco Cova, Christopher Kruegel, and Giovanni Vigna. "Peering through the iframe." In: *INFOCOM*. 2011. DOI: 10.1109/INFCOM.2011.5935193.

[27] Chris Grier, Andreas Pitsillidis, Niels Provos, M. Zubair Rafique, Moheeb Abu Rajab, Christian Rossow, Kurt Thomas, Vern Paxson, Stefan Savage, Geoffrey M Voelker, Lucas Ballard, Juan Caballero, Neha Chachra, Christian J Dietrich, Kirill Levchenko, Panayiotis Mavrommatis, Damon McCoy, and Antonio Nappa. "Manufacturing Compromise: The Emergence of Exploit-as-a-Service." In: *CCS*. New York, New York, USA: ACM Press, 2012, p. 821. DOI: 10.1145/2382196.2382283.

[28] Antonio Nappa, M. Zubair Rafique, and Juan Caballero. "The MALICIA dataset: identification and analysis of drive-by download operations." In: *International Journal of Information Security* 14.1 (2014). DOI: 10.1007/s10207-014-0248-7.

[29] Damon McCoy, A Pitsillidis, G Jordan, N Weaver, C Kreibich, B Krebs, G M Voelker, S Savage, and K Levchenko. "PharmaLeaks: Understanding the Business of Online Pharmaceutical Affiliate Programs." In: *USENIX Security* (2012), pp. 1–16.

[30] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. "Measuring and Analyzing Search-Redirection Attacks in the Illicit Online Prescription Drug Trade." In: *USENIX Security*. 2011.

[31] Nicolas Christin. "Traveling the silk road." In: *WWW*. New York, New York, USA: ACM Press, 2013, pp. 213–224. DOI: 10.1145/2488388.2488408.

[32] Rolf van Wegberg, Samaneh Tajalizadehkhoob, Kyle Soska, Ugur Akyazi, Carlos Hernandez Ganan, Bram Klievink, Nicolas Christin, and Michel van Eeten. "Plug and Prey? Measuring the Commoditization of Cybercrime via Online Anonymous Markets." In: *USENIX Security*. 2018, pp. 1009–1026.

[33] Manos Antonakakis, Tim April, Michael Bailey, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J Alex Halderman, Damian Menscher, Chad Seaman, Nick Sullivan, Kurt Thomas, Yi Zhou, Manos Antonakakis Tim April, Matthew Bernhard Elie Bursztein, Jaime J Cochran Zakir Durumeric Alex Halderman Luca Invernizzi, Michalis Kallitsis, Deepak Kumar, Chaz Lever Zane Ma, Joshua Mason, and Nick Sullivan Kurt Thomas. "Understanding the Mirai Botnet." In: *USENIX Security*. 2017.

[34] Christian Rossow, Christian Dietrich, and Herbert Bos. "Large-Scale Analysis of Malware Downloaders." In: *DIMVA*. Vol. 7591. 2013, pp. 42–61. DOI: 10.1007/978-3-642-37300-8_3.

[35] Oleksii Starov, Johannes Dahse, Syed Sharique Ahmad, Thorsten Holz, and Nick Nikiforakis. "No Honor Among Thieves: A Large-Scale Analysis of Malicious Web Shells." In: *WWW* (2016), pp. 1021–1032. DOI: 10.1145/2872427.2882992.

[36] Brett Stone-Gross, Christopher Kruegel, Kevin Almeroth, Andreas Moser, and Engin Kirda. "FIRE: FInding Rogue nEtworks." In: *ACSAC*. 2009, pp. 231–240. DOI: 10.1109/ACSAC.2009.29.

[37] Sumayah Alrwais, Xiaojing Liao, Xianghang Mi, Peng Wang, XiaoFeng Wang, Feng Qian, Raheem Beyah, and Damon McCoy. "Under the Shadow of Sunshine : Understanding and Detecting Bulletproof Hosting on Legitimate Service Provider Networks." In: *S&P*. 2017.

[38] Giovanni Sartor. *Providers Liability : From the eCommerce Directive to the future*. Tech. rep. EU Parliment Directorate General for Internal Policies (IP/A/IMCO/2017-07) (PE 614.179), 2017.

[39] Alice Hutchings, Richard Clayton, and Ross Anderson. "Taking down websites to prevent crime." In: *eCrime*. IEEE, 2016. DOI: 10.1109/ECRIME.2016.7487947.

[40] Stopbadware. *Web Hosting Provider Liability for Malicious content*. Tech. rep. 2011.

[41] European Commission. *Digital Single Market Policy on Illegal content on online platforms*. 2018. URL: https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms.

[42] OECD. *The Role of Internet Intermediaries in Advancing Public Policy Objectives*. OECD Publishing, 2011. DOI: 10.1787/9789264115644-en.

[43] Huw Fryer, Sophie Stalla-Bourdillon, and Tim Chown. "Malicious web pages: What if hosting providers could actually do something." In: *Computer Law & Security Review* 31.4 (2015). DOI: 10.1016/j.clsr.2015.05.011.

[44] Samaneh Tajalizadehkhoob, Maciej Korczynski, Arman Noroozian, Carlos Ganan, and Michel van Eeten. "Apples, oranges and hosting providers: Heterogeneity and security in the hosting market." In: *NOMS*. IEEE, 2016. DOI: 10.1109/NOMS.2016.7502824.

[45] Wouter de Vries. *Hosting provider Antagonist automatically fixes vulnerabilities in customers' websites*. 2012. URL: https://www.antagonist.nl/blog/2012/11/hosting-provider-antagonist-automatically-fixes-vulnerabilities-in-customers-websites/.

[46] Antonio Nappa, Richard Johnson, Leyla Bilge, Juan Caballero, and Tudor Dumitras. "The Attack of the Clones: A Study of the Impact of Shared Code on Vulnerability Patching." In: *IEEE S&P*. 2015, pp. 692–708. DOI: 10.1109/SP.2015.48.

[47] M3AAWG. *Anti-Abuse Best Common Practices for Hosting and Cloud Service Providers*. Tech. rep. March. Message, Mobile and Malware Anti-Abuse Working Group, 2015.

[48] Huw Fryer, Roksana Moore, and Tim Chown. "On the Viability of Using Liability to Incentivise Internet Security." In: *WEIS*. 2008. 2013, pp. 1–22.

[49] Ross Anderson, Rainer Bohme, Richard Clayton, and Tyler Moore. *Security Economics and The Internal Market*. Tech. rep. ENISA (European Network and Information Security Agency), 2008.

[50] Tim Berners-Lee. *30 years on, what's next #ForTheWeb?* 2019. URL: https://webfoundation.org/2019/03/web-birthday-30/.

[51] Davide Canali, Davide Balzarotti, and Aurélien Francillon. "The role of web hosting providers in detecting compromised websites." In: *WWW*. 2013.

[52] Danny Bradbury. "Testing the defences of bulletproof hosting companies." In: *Network Security* 2014.6 (2014). DOI: 10.1016/S1353-4858(14)70059-5.

[53] Mohammad Hanif Jhaveri, Orcun Cetin, Carlos Gañán, Tyler Moore, and Michel Van Eeten. "Abuse Reporting and the Fight Against Cybercrime." In: *ACM Computing Surveys* 49.4 (2017). DOI: 10.1145/3003147.

[54] Orcun Cetin, Carlos Gañán, Maciej Korczyski, and Michel Van Eeten. "Make Notifications Great Again: Learning How to Notify in the Age of Large-Scale Vulnerability Scanning." In: *WEIS*. 2017.

[55] Ben Stock, Giancarlo Pellegrino, Frank Li, Michael Backes, and Christian Rossow. "Didn't You Hear Me ? - Towards More Successful Web Vulnerability Notifications." In: *NDSS*. 2018.

[56] Samuel Marchal, Jérôme François, Radu State, and Thomas Engel. "Proactive discovery of phishing related domain names." In: *RAID*. Vol. 7462. LNCS. 2012, pp. 190–209. DOI: 10.1007/978-3-642-33338-5.

[57] Shuang Hao, Alex Kantchelian, Brad Miller, Vern Paxson, and Nick Feamster. "PREDATOR : Proactive Recognition and Elimination of Domain Abuse at Time-Of-Registration." In: *CCS*. 2016. DOI: 10.1145/2976749.2978317.

[58] Mahmood Sharif, Jumpei Urakawa, Nicolas Christin, Ayumu Kubota, and Akira Yamada. "Predicting Impending Exposure to Malicious Content from User Behavior." In: *CCS*. 2018, pp. 1487–1501. DOI: 10.1145/3243734.3243779.

[59] Joe Deblasio, Stefan Savage, Geoffrey M Voelker, and Alex C Snoeren. "Tripwire: Inferring Internet Site Compromise." In: *IMC*. 2017. DOI: 10.1145/3131365.3131391.

[60] Tyler Moore and Richard Clayton. "Examining the impact of website take-down on phishing." In: *eCrime*. ACM Press, 2007. DOI: 10.1145/1299015.1299016.

[61] Neha Chachra, Damon Mccoy, Stefan Savage, and Geoffrey M Voelker. "Empirically Characterizing Domain Abuse and the Revenue Impact of Blacklisting." In: *WEIS*. 2014.

[62] Brian Krebs. *DDoS-for-Hire Service Webstresser Dismantled*. 2018. URL: https://krebsonsecurity.com/2018/04/ddos-for-hire-service-webstresser-dismantled/.

[63] Charlie Osborne. *MaxiDed, dead: Law enforcement closes hosting service linked to criminal activity*. 2018. URL: https://www.zdnet.com/article/maxided-dead-law-enforcement-takes-down-bulletproof-hosting-linked-to-criminal-activity/.

[64] David Y Wang, Matthew Der Mohammad, Lawrence Saul, Damon Mccoy, Stefan Savage, and Geoffrey M Voelker. "Search + Seizure : The Effectiveness of Interventions on SEO Campaigns." In: *IMC*. 2014. DOI: 10.1145/2663716.2663738.

[65] Ross Anderson. "Why information security is hard - an economic perspective." In: *ACSAC*. 2001. DOI: 10.1109/ACSAC.2001.991552.

[66] Tyler Moore, Richard Clayton, and Ross Anderson. "The economics of online crime." In: *The Journal of Economic Perspectives* 23.3 (2009).

[67] T Moore and R Clayton. "The impact of incentives on notice and take-down." In: *WEIS*. 2009.

[68] Carl Shapiro and Hal R. Varian. *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Review Press, 1998.

[69] L Jean Camp and Catherine Wolfram. "Pricing Security." In: *Economics of Information Security*. Springer, 2004. Chap. 2, pp. 17–34. DOI: 10.1007/1-4020-8090-5_2.

[70] Nektarios Leontiadis. "Structuring Disincentives for Online Criminals." PhD thesis. Carnegie Mellon University, 2014.

[71] Ross Anderson, Chris Barton, Bohme Rainer, Richard Clayton, Michel van Eeten, Michael Levi, Tyler Moore, and Stefan Savage. "Measuring the Cost of Cybercrime." In: *WEIS*. 2012.

[72] Richard Clayton and Tony Mansfield. "A study of Whois privacy and proxy service abuse." In: *WEIS*. 2014.

[73] Yang Liu, Armin Sarabi, Jing Zhang, Parinaz Naghizadeh, Manish Karir, Michael Bailey, and Mingyan Liu. "Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents." In: *USENIX Security*. 2015.

[74] LegitScript and KnujOn. *Rogues and Registrars Report*. Tech. rep. LegitScript.com, 2010. URL: http://www.legitscript.com/download/Rogues-and-Registrars-Report.pdf.

[75] Nektarios Leontiadis and Nicolas Christin. "Empirically measuring WHOIS misuse." In: *ESORICS*. 2014.

[76] Matheus Xavier Ferreira, S. Matthew Weinberg, Danny Yuxing Huang, Nick Feamster, and Tithi Chattopadhyay. "Selling a Single Item with Negative Externalities." In: *WWW*. ACM Press, 2019. DOI: 10.1145/3308558.3313692.

[77] Zakir Durumeric, James Kasten, David Adrian, J. Alex Halderman, Michael Bailey, Frank Li, Nicolas Weaver, Johanna Amann, Jethro Beekman, Mathias Payer, and Vern Paxson. "The Matter of Heartbleed." In: *IMC*. 2014. DOI: 10.1145/2663716.2663755.

[78] Samaneh Tajalizadehkhoob, Tom van Goethem, Maciej Korczyński, Arman Noroozian, Rainer Böhme, Tyler Moore, Wouter Joosen, and Michel van Eeten. "Herding Vulnerable Cats: A Statistical Approach to Disentangle Joint Responsibility for Web Security in Shared Hosting." In: *CCS*. 2017. DOI: 10.1145/3133956.3133971.

[79] Tom Van Goethem, Ping Chen, and Nick Nikiforakis. "Large-Scale Security Analysis of the Web : Challenges and Findings." In: *Trust and Trustworthy Computing*. 2014, pp. 110–126. DOI: 10.1007/978-3-319-08593-7_8.

[80] Michel van Eeten, Johannes M. Bauer, Hadi Asghari, Shirin Tabatabaie, and Dave Rand. "The role of internet service providers in botnet mitigation an empirical analysis based on spam data." In: *WEIS*. 2010.

[81] Michel van Eeten, Hadi Asghari, Johannes M. Bauer, and Shirin Tabatabaie. *Internet Service Providers and Botnet Mitigation (A Fact-Finding Study on the Dutch Market)*. Tech. rep. Delft University of Technology, 2011.

[82] Hadi Asghari, Michel J.G. van Eeten, and Johannes M. Bauer. "Economics of Fighting Botnets: Lessons from a Decade of Mitigation." In: *S&P* 13.5 (2015), pp. 16–23. DOI: 10.1109/MSP.2015.110.

[83] Ross Anderson and Tyler Moore. "The Economics of Information Security." In: *Science* 314.5799 (2006), pp. 610–613. DOI: 10.1126/science.1130992.

[84] Nicolas Christin, Sally S Yanagihara, and Keisuke Kamataki. "Dissecting One Click Frauds." In: *ACM CCS*. 2010, pp. 15–26.

[85] J Zhang, Z Durumeric, and M Bailey. "On the Mismanagement and Maliciousness of Networks." In: *NDSS*. 2014. DOI: 10.14722/ndss.2014.23057.

[86] Daniel E Geer. "Less Is More : Saving the Internet from Itself." In: *IEEE S&P Magazine* 13.1 (2015), p. 80.

[87] Dinei Florencio and Cormac Herley. "Where Do All the Attacks Go?" In: *Economics of Information Security and Privacy III*. 2013, pp. 13–33. DOI: 10.1007/978-1-4614-1981-5.

[88] Cormac Herley. "Security, cybercrime, and scale." In: *CACM* 57.9 (2014), pp. 64–71. DOI: 10.1145/2654847.

[89] Marie Vasek and Tyler Moore. "Identifying Risk Factors for Webserver Compromise." In: *Financial Cryptography and Data Security*. Vol. 8437. LNCS. 2014, pp. 326–345. DOI: 10.1007/978-3-662-45472-5.

[90] Marie Vasek, John Wadleigh, and Tyler Moore. "Hacking Is Not Random: A Case-Control Study of Webserver-Compromise Risk." In: *IEEE Transactions on Dependable and Secure Computing* 13.2 (2016), pp. 206–219. DOI: 10.1109/TDSC.2015.2427847.

[91] Richard Clayton, Tyler Moore, and Nicolas Christin. "Concentrating Correctly on Cybercrime Concentration." In: *WEIS*. 2015.

[92] Craig A. Shue, Andrew J. Kalafut, and Minaxi Gupta. "Abnormally Malicious Autonomous Systems and Their Internet Connectivity." In: *IEEE/ACM TON* 20.1 (2012), pp. 220–230. DOI: 10.1109/TNET.2011.2157699.

[93] Maria Konte, Roberto Perdisci, and Nick Feamster. "ASwatch: An AS Reputation System to Expose Bulletproof Hosting ASes." In: *SIGCOMM*. ACM Press, 2015. DOI: 10.1145/2785956.2787494.

[94] Cormac Herley and P. C.Van Oorschot. "SoK: Science, Security and the Elusive Goal of Security as a Scientific Pursuit." In: *IEEE S&P*. 2017. DOI: 10.1109/SP.2017.38.

[95] *HostExploit*. URL: http://hostexploit.com/.

[96] Shari Lawrence Pfleeger and Robert K. Cunningham. "Why measuring security is hard." In: *IEEE Security & Privacy Magazine* 8.4 (2010), pp. 46–54. DOI: 10.1109/MSP.2010.60.

[97] Samaneh Tajalizadehkhoob, Rainer Böhme, Carlos Gañán, Maciej Korczyński, and Michel Van Eeten. "Rotten Apples or Bad Harvest? What We Are Measuring When We Are Measuring Abuse." In: *CoRR* (2017).

[98] Netcraft. *Hosting Provider Analysis*. 2017. URL: https://www.netcraft.com/internet-data-mining/hosting-analysis/.

[99] Xue Cai, John Heidemann, Balachander Krishnamurthy, and Walter Willinger. "Towards an AS-to-organization map." In: *IMC*. 2010. DOI: 10.1145/1879141.1879166.

[100] Arman Noroozian, Maciej Korczynski, Samaneh Tajalizadehkhoob, and Michel van Eeten. "Developing Security Reputation Metrics for Hosting Providers." In: *USENIX CSET*. 2015.

[101] Arman Noroozian, Michael Ciere, Maciej Korczynski, Samaneh Tajalizadehkhoob, and Michel Van Eeten. "Inferring the Security Performance of Providers from Noisy and Heterogenous Abuse Datasets." In: *WEIS*. 2017.

[102] Arman Noroozian, Geoffrey Simpson, Maciej Korczynski, Tyler Moore, Rainer Bohme, and Michel van Eeten. "Using Abuse Data to Evaluate Remediation Efforts." 2018.

[103] Arman Noroozian, Jan Koenders, Eelco van Veldhuizen, Carlos Hernandez Ganan, Sumayah Alrwais, Damon McCoy, and Michel van Eeten. "Platforms in Everything: Analyzing Ground-Truth Data on the Anatomy and Economics of Bullet Proof Hosting." In: *Proc. of Usenix Security Symposium*. 2019.

[104] Arman Noroozian, Maciej Korczyński, Carlos Hernandez Gañan, Daisuke Makita, Katsunari Yoshioka, and Michel Van Eeten. "Who Gets the Boot? Analyzing Victimization by DDoS-as-a-Service." In: *Proc. of RAID*. 2016. DOI: 10.1007/978-3-319-45719-2_17.

[105] C. Wagner, J. François, R. State, A. Dulaunoy, T. Engel, and G. Massen. "ASMATRA: Ranking ASs providing transit service to malware hosters." In: *Integrated Network Management*. 2013, pp. 260–268.

[106] Zachary Weinberg, Shinyoung Cho, Nicolas Christin, Vyas Sekar, and Phillipa Gill. "How to catch when proxies lie :Verifying the physical locations of network proxies with active geolocation." In: *In ACM SIGCOMM IMC*. 2018, pp. 203–217. DOI: 10.1145/3278532.3278551.

[107] *Maxmind GeoIP2 DB*. URL: https://www.maxmind.com/en/geoip2-isp-database.

[108] Andreas Pitsillidis, Chris Kanich, Geoffrey M. Voelker, Kirill Levchenko, and Stefan Savage. "Taster's choice: a comparative analysis of spam feeds." In: *IMC*. 2012, p. 427. DOI: 10.1145/2398776.2398821.

[109] Leigh Metcalf and Jonathan M Spring. *Everything You Wanted to Know About Blacklists But Were Afraid to Ask*. Tech. rep. CERT Network Situational Awareness Group, 2013.

[110] Benjamin Zi Hao Zhao, Muhammad Ikram, Hassan Jameel Asghar, Mohamed Ali Kaafar, Abdelberi Chaabane, and Kanchana Thilakarathna. "A decade of mal-activity reporting: A retrospective analysis of internet malicious activity blacklists." In: *AsiaCCS*. 2019, pp. 193–205. DOI: 10.1145/3321705.3329834.

[111] Farsight Security. *DNSDB*. URL: https://www.dnsdb.info.

[112] M Korczynski, S Tajalizadehkhoob, A Noroozian, M Wullink, C Hesselman, and M Van Eeten. "Reputation Metrics Design to Improve Intermediary Incentives for Security of TLDs." In: *Euro S&P*. 2016.

[113] Lorrie Faith Cranor. "Declared-Strategy Voting: An Instrument for Group Decision-Making." PhD thesis. Washington University, 1996.

[114] Marc Kührer, Christian Rossow, and Thorsten Holz. "Paint It Black: Evaluating the Effectiveness of Malware Blacklists." In: *RAID*. Vol. 7462. LNCS. Cham, 2012, pp. 1–21. DOI: 10.1007/978-3-319-11379-1.

[115] Stjepan Gros, Mislav Stublić, and Leonardo Jelenović. "Determining autonomous systems reputation based on DNS measurements." In: *MIPRO*. 2012, pp. 1526–1528.

[116] Giovane C M Moura, Ramin Sadre, and Aiko Pras. "Bad neighborhoods on the internet." In: *IEEE Communications Magazine* 52.7 (2014), pp. 132–139. DOI: 10.1109/MCOM.2014.6852094.

[117] Hadi Asghari, Michael Ciere, and Michel J G Van Eeten. "Post-Mortem of a Zombie: Conficker Cleanup After Six Years." In: *USENIX Security*. 2015.

[118] Fanny Lalonde Levesque, Jose M. Fernandez, Anil Somayaji, and Dennis Batchelder. "National-level risk assessment : A multi-country study of malware infections." In: *WEIS*. 2016, pp. 1–30.

[119] Shu He, Gene Moo Lee, Sukjin Han, and Andrew B. Whinston. "How would information disclosure influence organizations' outbound spam volume? Evidence from a field experiment." In: *Journal of Cybersecurity* 2.1 (2016), pp. 99–118. DOI: 10.1093/cybsec/tyw011.

[120] Andrew J Kalafut, Craig A Shue, and Minaxi Gupta. "Malicious Hubs: Detecting Abnormally Malicious Autonomous Systems." In: *INFOCOM*. IEEE, 2010, pp. 1–5. DOI: 10.1109/INFCOM.2010.5462220.

[121] Benjamin Edwards, Steven Hofmeyr, Stephanie Forrest, and Michel van Eeten. "Analyzing and Modeling Longitudinal Security Data: Promise and Pitfalls." In: *ACSAC*. ACM Press, 2015, pp. 391–400. DOI: 10.1145/2818000.2818010.

[122] *StopBadware*. URL: https://www.stopbadware.org/data-sharing.

[123] *APWG*. URL: https://www.antiphishing.org/.

[124] *Phishtank*. URL: https://www.phishtank.com/index.php.

[125] Brian Greenhill, Michael D. Ward, and Audrey Sacks. "The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models." In: *American Journal of Political Science* 55.4 (2011), pp. 991–1002. DOI: 10.1111/j.1540-5907.2011.00525.x.

[126] Johannes Hartig and Jana Höhler. "Multidimensional IRT models for the assessment of competencies." In: *Studies in Educational Evaluation* 35.2-3 (2009), pp. 57–63. DOI: 10.1016/j.stueduc.2009.10.002.

[127] Lihua Yao. "Reporting Valid and Reliable Overall Scores and Domain Scores." In: *Journal of Educational Measurement* 47.3 (2016), pp. 339–360.

[128] Wolter Pieters, Sanne H.G. van der Ven, and Christian W. Probst. "A move in the security measurement stalemate." In: *Proceedings of the 2012 workshop on New security paradigms - NSPW '12*. ACM Press, 2012, pp. 1–14. DOI: 10.1145/2413296.2413298.

[129] W.R. Gilks, S. Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in Practice*. 1st ed. Chapman & Hall, 1996.

[130] *mc-stan*. URL: http://mc-stan.org/.

[131] Andrew Gelman and Donald B Rubin. "Inference from Iterative Simulation Using Multiple Sequences." In: *Statistical Science* 7.4 (1992), pp. 457–472.

[132] Benjamin Edwards, Jay Jacobs, and Stephanie Forrest. "Risky Business: Assessing Security with External Measurements." 2016.

[133] Kyle Soska and Nicolas Christin. "Automatically detecting vulnerable websites before they turn malicious." In: *USENIX Security*. 2014.

[134] *SpamHaus DBL*. URL: https://www.spamhaus.org/dbl/.

[135] Frank Li, Grant Ho, Eric Kuan, Yuan Niu, Lucas Ballard, Kurt Thomas, Elie Bursztein, and Vern Paxson. "Remedying Web Hijacking: Notification Effectiveness and Webmaster Comprehension." In: *WWW*. 2016. DOI: 10.1145/2872427.2883039.

[136] Marc Kührer, Thomas Hupperich, Christian Rossow, and Thorsten Holz. "Exit from Hell ? Reducing the Impact of Amplification DDoS Attacks." In: *USENIX Security*. 2014.

[137] CommTouch. *Compromised websites: an owner's perspective*. Tech. rep. February. 2012, 15pp.

[138] Niels Provos, Panayiotis Mavrommatis, Moheeb Abu Rajab, and Fabian Monrose. "All Your iFRAMEs Point to Us." In: *USENIX Security*. 2008.

[139] Christian Rossow. "Amplification Hell: Revisiting Network Protocols for DDoS Abuse." In: *NDSS*. 2014.

[140] Frank Li, Zakir Durumeric, Jakub Czyz, Mohammad Karami, Michael Bailey, Urbana Champaign, Damon Mccoy, New York, Stefan Savage, and Vern Paxson. "You ' ve Got Vulnerability : Exploring Effective Vulnerability Notifications." In: *USENIX Security*. 2016.

[141] Marie Vasek, Matthew Weeden, and Tyler Moore. "Measuring the Impact of Sharing Abuse Data with Web Hosting Providers." In: *ACM WISCS*. 2016. DOI: 10.1145/2994539.2994548.

[142] Kurt Thomas, Juan Antonio Elices Crespo, Ryan Rasti, Jean-Michel Picod, Damon Mccoy, Lucas Ballard, Elie Bursztein, Moheeb Abu Rajab, and Niels Provos. "Investigating Commercial Pay-Per-Install and the Distribution of Unwanted Software." In: *USENIX Security*. 2016.

[143] Dhia Mahjoub and Sarah Brown. *Behaviors and Patterns of Bulletproof and Anonymous Hosting Providers*. 2017. URL: https://www.usenix.org/conference/enigma2017/conference-program/presentation/mahjoub.

[144] Xiao Han, Nizar Kheir, and Davide Balzarotti. "The Role of Cloud Services in Malicious Software: Trends and Insights." In: *DIMVA*. 2015. DOI: 10.1007/978-3-319-20550-2_10.

[145] Carlos Gañán, Orcun Cetin, and Michel van Eeten. "An Empirical Analysis of ZeuS C&C Lifetime." In: *ASIACCS* (2015). DOI: 10.1145/2714576.2714579.

[146] *Webmaster Response*. URL: https://transparencyreport.google.com/safe-browsing/overview.

[147] Mozilla. *Public Suffix List*. URL: https://publicsuffix.org/.

[148] Dirk F. Moore. *Applied Survival Analysis Using R*. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-31245-3.

[149] Jakub Czyz, Michael Kallitsis, Christos Papadopoulos, and Michael Bailey. "Taming the 800 Pound Gorilla : The Rise and Decline of NTP DDoS Attacks." In: *IMC*. 2014.

[150] Brian Krebs. *Inside the Gozi Bulletproof Hosting Facility*. 2013. URL: https://krebsonsecurity.com/2013/01/inside-the-gozi-bulletproof-hosting-facility/.

[151] Brian Krebs. *Host of Internet Spam Groups Is Cut Off*. 2008. URL: http://www.washingtonpost.com/wp-dyn/content/article/2008/11/12/AR2008111200658.html.

[152] Brian Krebs. *Shadowy Russian Firm Seen as Conduit for Cybercrime*. 2007. URL: http://www.washingtonpost.com/wp-dyn/content/article/2007/10/12/AR2007101202461.html.

[153] Patrick Howell O'Neill. *An in-depth guide to Freedom Hosting, the engine of the Dark Net*. 2013. URL: https://www.dailydot.com/news/eric-marques-tor-freedom-hosting-child-porn-arrest/.

[154] Dutch-Police. *Nederlandse en Thaise politie pakken bulletproof hoster aan*. URL: https://www.politie.nl/nieuws/2018/mei/16/11-nederlandse-en-thaise-politie-pakken-bulletproof-hoster-aan.html.

[155] Catalin Cimpanu. *Police Seize Servers of Bulletproof Provider Known For Hosting Malware Ops*. URL: https://www.bleepingcomputer.com/news/security/police-seize-servers-of-bulletproof-provider-known-for-hosting-malware-ops/ (visited on 05/28/2019).

[156] Shuang Hao, Kevin Borgolte, Nick Nikiforakis, Gianluca Stringhini, Manuel Egele, Michael Eubanks, Brian Krebs, and Giovanni Vigna. "Drops for Stuff: An Analysis of Reshipping Mule Scams." In: *CCS*. 2015, pp. 1081–1092. DOI: 10.1145/2810103.2813620.

[157] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. "The Menlo report." In: *IEEE Security and Privacy* 10.2 (2012), pp. 71–75. DOI: 10.1109/MSP.2012.52.

[158] Annelie Langerak. *Groot pedonetwerk opgerold*. 2018. URL: https://www.telegraaf.nl/nieuws/2043709/groot-pedonetwerk-opgerold.

[159] Damon Mccoy, Hitesh Dharmdasani, Christian Kreibich, Geoffrey M Voelker, and Stefan Savage. "Priceless : The Role of Payments in Abuse-advertised Goods." In: *CCS* (2012), pp. 845–856. DOI: 10.1145/2382196.2382285.

[160] Andy Greenberg. *Operation Bayonet: Inside the Sting That Hijacked an Entire Dark Web Drug Market*. URL: https://www.wired.com/story/hansa-dutch-police-sting-operation/ (visited on 11/01/2018).

[161] *CleanMX*. URL: https://support.clean-mx.com.

[162] Brett Stone-Gross, Marco Cova, Lorenzo Cavallaro, Brett S Gross, Marco Cova, Lorenzo Cavallaro, Bob Gilbert, Martin Szydlowski, Richard Kemmerer, Christopher Kruegel, Giovanni Vigna, Brett Stone-gross, Marco Cova, Lorenzo Cavallaro, Bob Gilbert, Martin Szydlowski, Richard Kemmerer, Christopher Kruegel, and Giovanni Vigna. "Your Botnet is My Botnet : Analysis of a Botnet Takeover." In: *CCS*. Vol. 97. 3. 2009, pp. 635–647. DOI: 10.1145/1653662.1653738.

[163] Sumayah Alrwais, Kan Yuan, Eihal Alowaisheq, Zhou Li, and Xiaofeng Wang. "Understanding the Dark Side of Domain Parking." In: *USENIX Security*. 2014.

[164] Kyle Soska and Nicolas Christin. "Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem." In: *USENIX Security*. 2015, pp. 33–48. DOI: 10.1007/s00253-017-8456-5.

[165] Ryan Brunt, Prakhar Pandey, and Damon McCoy. "Booted: An Analysis of a Payment Intervention on a DDoS-for-Hire Service." In: *WEIS* (2017).

[166] Lukas Kramer, Johannes Krupp, Daisuke Makita, Tomomi Nishizoe, Takashi Koide, Katsunari Yoshioka, and Christian Rossow. "AmpPot : Monitoring and Defending Against Amplification DDoS Attacks." In: *RAID*. 2015.

[167] José Jair Santanna and Anna Sperotto. "Characterizing and mitigating the DDoS-as-a-Service phenomenon." In: *AIMS*. 2014, pp. 74–78. DOI: 10.1007/978-3-662-43862-6_10.

[168] Marc Kührer, Thomas Hupperich, Jonas Bushart, Christian Rossow, and Thorsten Holz. "Going Wild : Large-Scale Classification of Open DNS Resolvers Categories and Subject Descriptors." In: *IMC*. 2015.

[169] Mohammad Karami and Damon Mccoy. "Understanding the Emerging Threat of DDoS-As-a-Service." In: *USENIX LEET*. 2013, pp. 2–5.

[170] Jose Jair Santanna, Romain Durban, Anna Sperotto, and Aiko Pras. "Inside Booters: An Analysis on Operational Databases." In: *IM*. 2015, pp. 432–440. DOI: 10.1109/INM.2015.7140320.

[171] Mohammad Karami, Youngsam Park, and Damon McCoy. "Stress testing the Booters: Understanding and undermining the business of DDoS services." In: *WWW*. 2016. DOI: http://dx.doi.org/10.1145/2872427.2883004.

[172] Akamai. *State of the Internet / Security Q4*. Tech. rep. Akamai, 2014. URL: https://www.stateoftheinternet.com/.

[173] Arbor Networks. *Worldwide Infrastructure Security Report Volume X*. Tech. rep. 2015. DOI: 10.1016/S1353-4858(97)83501-5. URL: https://www.arbornetworks.com/insight-into-the-global-threat-landscape.

[174] Incapsula. *DDoS Global Threat Landscape Report*. Tech. rep. 2015. URL: http://lp.incapsula.com/ddos-report-2015.html.

[175] Jair Santanna, Roland Van Rijswijk-deij, Rick Hofstede, and Anna Sperotto. "Booters - An Analysis of DDoS-as-a-Service Attacks." In: *IM*. 2015.

[176] Kaspersky. *Statistics on Botnet Assisted DDoS Attacks*. 2015. URL: https://securelist.com/blog/research/70071/statistics-on-botnet-assisted-ddos-attacks-in-q1-2015/.

[177] TeleGeography. *Telegeography Globalcomms Data*. URL: http://shop.telegeography.com/products/globalcomms-database.

[178] CAIDA. *AS Classification*. URL: http://www.caida.org/data/as-classification/.

[179] Akamai. *State of the Internet / Security Q4*. Tech. rep. 2015. URL: https://www.stateoftheinternet.com/downloads/pdfs/q4-2015-security-report-ddos-stats-trends-analysis-infographic.pdf.

[180] PRB. *Population Reference Bureau - Gross Domestic Product*. URL: http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=260.

[181] Andrew M. Ledbetter and Jeffrey H. Kuznekoff. "More Than a Game: Friendship Relational Maintenance and Attitudes Toward Xbox LIVE Communication." In: *Communication Research* 39.2 (2012), pp. 269–290. DOI: 10.1177/0093650210397042.

[182] Miltiadis Allamanis, Salvatore Scellato, and Cecilia Mascolo. "Evolution of a location-based online social network." In: *IMC*. New York, New York, USA: ACM Press, 2012, p. 145. DOI: 10.1145/2398776.2398793.

[183] Freek Schravese and Arjen Born. *Lekker thuis providers platleggen*. 2015. URL: http://www.nrc.nl/handelsblad/2015/10/17/lekker-thuis-providers-platleggen-1545974.

[184] Alexa. *Alexa Top 1M Ranked Sites*. 2015. URL: http://s3.amazonaws.com/alexa-static/top-1m.csv.zip.

[185] Jonathan Zittrain, Kendra Albert, and Lawrence Lessig. "Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations." In: *Legal Information Management* 14.02 (2014), pp. 88–99. DOI: http://dx.doi.org/10.1017/S1472669614000255.

[186] E. L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations." In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481.

[187] Marc Kührer, Thomas Hupperich, Christian Rossow, Thorsten Holz, and G Horst. "Hell of a Handshake : Abusing TCP for Reflective Amplification DDoS Attacks." In: *USENIX WOOT*. 2014.

[188] Zakir Durumeric, Michael Bailey, and J Alex Halderman. "An Internet-Wide View of Internet-Wide Scanning." In: *USENIX Security*. 2014, pp. 65–78.

[189] Alice Hutchings and Richard Clayton. "Exploring the Provision of Online Booter Services." In: *Deviant Behavior* (2016), pp. 1–16. DOI: 10.1080/01639625.2016.1169829.

[190] Dinei Florencio and Cormac Herley. "Is Everything We Know About Password-Stealing Wrong?" In: *IEEE Security & Privacy Magazine* 10.6 (2012). DOI: 10.1109/MSP.2012.57.

[191] Walter W. Powell. "Neither Market nor Hierarchy: Network Forms of Organization." In: *Research in Organizational Behavior* (1990). DOI: 10.1007/978-3-658-21742-6_108.

[192] Anne Mette Kjaer. *Governance*. Polity Press, 2004.

[193] Tim Tenbensel. "Multiple modes of governance." In: *Public Management Review* 7.2 (2005), pp. 267–288. DOI: 10.1080/14719030500091566.

[194] Michel van Eeten. "Patching security governance: an empirical view of emergent governance mechanisms for cybersecurity." In: *Digital Policy, Regulation and Governance* 19.6 (2017), pp. 429–448. DOI: 10.1108/DPRG-05-2017-0029.

[195] Suqi Liu, Ian Foster, Stefan Savage, Geoffrey M Voelker, and Lawrence K Saul. "Who is .com?: Learning to Parse WHOIS Records." In: *IMC* (2015), pp. 369–380. DOI: 10.1145/2815675.2815693.

[196] Laurence Dodds. *British MPs call for German-style law to block hate speech on social media*. 2018. URL: https://www.telegraph.co.uk/technology/2018/07/28/british-mps-call-german-style-law-block-hate-speech-social-media/.

[197] *France online hate speech law to force social media sites to act quickly*. 2019. URL: https://www.theguardian.com/world/2019/jul/09/france-online-hate-speech-law-social-media.

[198] European-Commission. *Code of Conduct on countering illegal hate speech online: Questions and answers on the fourth evaluation*. 2019. URL: http://europa.eu/rapid/press-release{_}MEMO-19-806{_}en.htm.

[199] Jillian C. York. *European Commission's Hate Speech Deal With Companies Will Chill Speech*. URL: https://www.eff.org/deeplinks/2016/06/european-commissions-hate-speech-deal-companies-will-chill-speech (visited on 07/19/2019).

SUMMARY

Cloud, hosting, and Internet Service Providers (ISPs), as well as social media platforms, have made it very easy to for individuals to create and place content online through their services. The streamlined services of such so-called Internet intermediaries have also enabled miscreants to misuse the Internet as a platform for cybercrime and illicit financial gain.

Among the many Internet intermediary firms, hosting providers are entities whose services are highly prone to abuse by miscreants. Their services, which typically include the provisioning of servers, Internet connectivity, and online storage capacity, function as the backbone of many forms of cybercrime and have been shown to be abused regularly. For instance, miscreants commonly attempt to steal banking details, account credentials, or other forms of sensitive user data via hosted phishing pages. Or they extort ransom money by spreading hosted ransomware binaries or by launching distributed denial of service attacks against online institutions via hosted command-and-control infrastructure of botnets.

Such forms of cybercrime not only affect individuals, but also businesses globally, as well as society as a whole. Therefore an increasingly important question discussed among academics and policy makers is one concerning the role that hosting providers can (or should) play to prevent the abuse of their services. Hosting providers are theoretically in a key position to combat cybercrime as they are often the entities renting out the abused resources. Yet, notwithstanding providers' current security measures to combat abuse, their response varies widely. In many cases the response is lacking in effectiveness, as empirical evidence suggests.

At its core, improving security within a global hosting market constitutes a collective action problem. This means that it is unlikely for individual hosting providers to be able to address the problem alone. An abundance of insecure hosting providers would still allow miscreants to conduct their illicit activities online even if smaller groups of providers effectively prevent the abuse of their services. Moreover, this matter is also an incentives problem due to the costs associated with combating abuse within a highly competitive global market that typically operates under low profit margins. As a result, before solutions to the abuse problem can be devised, it is critical to address the incentives problem.

Within the global hosting market, the apparent lack of incentives to combat abuse is combined with information asymmetry about the security efforts and practices of hosting providers. We currently lack empirical information about key aspects of this market. For instance, we lack information as basic as the list of providers that operate within the market. We have inadequate techniques of identifying them, insufficient information on the providers that are abused more often than others, as well as a lack of information on the effectiveness of current security practices of hosting providers against abuse. This type of information asymmetry exacerbates the abuse problem and weakens provider incentives to combat the cybercrime facilitated by their services. And since we have no way of telling apart 'good' and 'bad' hosting providers, a predictable outcome where most providers do not effectively combat abuse. This is a phenomenon driven by the misalignment of their economic interests and the need for better security, i.e. their security incentives.

Theoretically, various governance schemes may be devised to realign provider security incentives in ways to steer the global hosting market towards a more desirable security outcome. Regulation through top-down hierarchical governance approaches may prove to be effective, e.g., by placing liability onto providers for the negative externalities of cybercrime facilitated through their services. This would undermine, however, safe harbor mechanisms that have allowed online services to innovate. Softer forms of regulation, for example market-based mechanisms to reward good providers and/or punish frequently abused ones may also steer the market in the right direction. Self-regulation and voluntary forms of action through network-based or community-driven efforts may also positively affect security in the hosting market albeit in limited ways. Nevertheless, I argue that independent of any particular governance scheme to realign provider security incentives, the critical problem that needs to be solved first, is that of information asymmetry, i.e. to overcome our knowledge shortcomings of this market and to devise methods by which to distinguish good and bad hosting providers. This leads to main question that this thesis attempts to answer, namely:

> How can we quantify the effectiveness of hosting provider security practices?

In summary, I propose to design and operationalize security metrics as a means for quantifying and comparing the effectiveness of hosting provider security practices. In each following chapter I breakdown and explore multiple paths of inquiry into the main research question that I pose above.

In Chapter 1 I first provide a more extensive discussion of the state of security in the global hosting market and the motivation behind and importance of the research question that I pose.

Next, Chapter 2 is concerned with developing metrics to compare the security efforts of hosting providers from empirical observations of abuse incidents. Here, I employ empirical data from so-called 'abuse feeds' and identify systematic steps to translate captured incident data into metrics reflective of hosting provider security postures. The aim is to produce metrics for comparing the security efforts of providers that are meaningful and stable. Two types of metrics are defined. First are a set of metrics based on how frequently abuse incidents occur. I argue that these are reflective of **proactive** provider security efforts since they signal how well security incidents are prevented from occurring in the first place. Second are a set of metrics based on how timely incidents are remediated by the providers which reflect their **reactive** security efforts once incidents have already occurred. Based on the abuse data, I find abuse incidents to be mostly concentrated around a few hosting providers. I also find that proactive and reactive security efforts to be only weakly correlated thus suggesting that each type of metric captures an independent dimension of provider security effort. This initial chapter then sets the agenda and serves as a road map for what subsequent steps to take to answer the main research question.

In Chapter 3 I develop improved techniques for answering the question of how the proactive dimension of provider security efforts can be externally measured and how the inherent noisy nature of the abuse data, on which much of my research relies, may be dealt with. Here I combine Bayesian statistical methods and Item Response Theory to estimate and compare the security efforts of hosting providers as a latent trait indirectly observable through abuse data. The main contributions here can be summarized as a set of estimates that more robustly represent the proactive dimension of provider security with associated confidence bounds, as well as demonstrating the power of these estimates to explain and predict the concentration of abuse incidents at the global hosting market scale. These results critically depend upon a theoretical causal model of abuse that formalizes how hosting provider characteristics, attacker behavior and security efforts moderate and affect the observed frequencies provider abuse.

Subsequently, in Chapter 4 I investigate how hosting providers react when incidents occur and how well they perform when notified of security incidents, i. e. compare their reactive security efforts. I approached this by collecting incident data and comparing the amount of time required to take action against discovered harmful content on hosting provider networks. While the approach yields a more direct measurement of security effort, I find that complexities arise from remediation times being influenced by actors other than the providers themselves, namely webmasters whose resources may have been compromised, attackers who may attempt to host harmful content directly themselves, as well as other stakeholders that may notify providers of

abuse through various means. Due to these additional complexities, I formalize the factors that may influence remediation times through an explanatory model which I then use to compare and explain the differences among provider reactive security efforts. The main contributions here may be summarized as a set of techniques to measure and compare remediation times among providers as well as methods to draw more robust inferences from the noisy abuse data.

Chapter 5 takes a closer look at the special case of criminal Bullet-Proof Hosting (BPH) providers. These, which are a difficult area of the hosting market to tackle, are hosting providers that knowingly allow miscreants to host harmful content through their services and even assist in its online persistence thereby enabling a large range of cybercrime. In particular I investigate how such providers operate and whether they can be identified through security performance metrics, e. g. as metric outliers. BPH providers are interesting cases to examine as they may demonstrate the limitations of security metrics. My findings here include a wider range of unique insights into how BPH providers operate internally, as well as insights regarding their economics and profitability or rather lack thereof. Here, I demonstrate that due to their modus operandi, more sophisticated 'agile' BPH providers may not be detectable through the security metrics that I have developed and they may be unsuitable tools to detect BPH.

In Chapter 6, I step back and examine the negative side-effects of provider security negligence by studying the victims of cybercrime in a case-study of Distributed Denial of Service (DDoS) attacks which are facilitated in part by negligent hosting providers that host 'booter' websites. Booter websites package and sell the ability to kick arbitrary targets offline by launching Distributed Denial of Service (DDoS) attacks at the click of a button. The empirical insights gained in this study demonstrate the extent of harm that malicious content may cause if not adequately addressed, particularly by hosting providers that host them and fail to to take them offline. This study reveals that the bulk of the targeted victims of DDoS attacks emanating from booter websites are regular Internet users, most notably online gamers. The observed pattern demonstrates how often such attacks take place and target regular Internet users. While security industry reports commonly list high profile businesses as victims of DDoS attacks, this study demonstrates that the consequences of negligence are much more wide spread and go beyond business and affect society as a whole.

Finally, in Chapter 7 I bring together my findings and results and discuss the implications of my findings along with my concluding remarks. I examine how security metrics may help in devising governance strategies to move the hosting market towards more desirable security outcomes and how they may be incentivized to take more effective steps against abuse.

AUTHORSHIP CONTRIBUTIONS

The dissertation is based on five peer-reviewed empirical studies that resulted from collaborative work with several co-authors. While all of these studies were lead by myself, I was fortunate enough to receive the valuable support and contributions of my co-authors in each of these studies. I will outline their contributions to each study below.

For the first study, discussed in Chapter 2, my co-authors, Samaneh Tajalizadehkhoob, Maciej Korczyński and my promoter Michel van Eeten, have helped with improving the draft, clarifying its arguments, proof reading and polishing of the text. Collection of the underlying data used for this study has mostly been a collaborative effort and my co-authors, particularly Dr. Maciej Korczyński has been instrumental in contacting and setting up data sharing agreements with some of the providers of the abuse data. Code to wrangle and analyze the abuse data has been jointly developed by myself and Maciej Korczyński. Code to wrangle passive DNS data from Farsight Security's DNSDB and extract information regarding the size of hosting providers was developed by myself. Identification of hosting providers, their ASNs and attribution of incidents to their networks was done through the use of the `pyasn` python package largely developed by Hadi Asghari.

For the second study, discussed in Chapter 3, my co-authors Michael Ciere, Maciej Korczyński, Samaneh Tajalizadehkhoob and Michel van Eeten, to varying degrees have greatly helped with improving my drafts and clarifying its arguments. Collection of the underlying data for this paper was greatly supported by the efforts of Maciej Korczyński and is indebted to the support of Marie Vasek and Tyler Moore for providing me with access to StopBadware Data. Code to wrangle data and extract information from the additional abuse feeds was developed by myself. Modelling of the data and development of the models in R and the `mc-stan` package received valuable support from Michael Ciere who originally suggested the use of Item Response Theory in this chapter. Additional code was also jointly developed by myself and Samaneh Tajalizadehkhoob to map the hosting market through WHOIS IP allocation, which was also used for a second co-authored study lead by Samaneh Tajalizadehkhoob.

For the third study which I lead, discussed in Chapter 4, my co-authors, Geoffrey Simpson, Maciej Korczyński, Tyler Moore, Rainer Böhme and my promoter Michel van Eeten, as before have greatly helped with improving text and clarifying its arguments. Code to

collect, model and analyze the data was developed by myself. Tyler Moore and Rainer Böhme provided valuable support in setting up the study, interpreting the results and triangulating its findings. They have both contributed to fleshing out the methodology needed to analyze the data and scoping of the study's central question. I also received valuable support from Geoffrey Simpson in collecting Google SafeBrowsing data through StopBadware.

For the forth study, discussed in Chapter 5, my co-authors Jan Koenders and Eelco van Veldhuizen, from the Dutch National High-Tech Crime Unit, including other members of their team, have provided the underlying data and helped with its interpretation, proof reading, and improving my text. Carlos Ganan, Sumayah Alrwais, Damon McCoy and Michel van Eeten have greatly help to polish the text, place it within the context of the related work and focus the main research question of the paper. Sumayah Alrwais also provided me with additional data on the landscape of BPH providers.

Finally my fifth study, discussed in Chapter 6, received greatly valued support from my co-authors, Maciej Korczynski, Carlos Ganan, Daisuke Makita, Katsunari Yoshioka, and Michel Van Eeten. They, have as before helped me with improving the text and focusing the research question. Daisuke Makita and Katsunari Yoshioka have graciously provided the AmpPot data for this paper. Carlos Ganan also guided and provided valuable support in modeling the data.

While I have lead these studies, I greatly appreciate the support and help of my co-authors. They have to varying degrees contributed to my studies in terms of ideas, feedback, writing, and most importantly their time. I would like to especially acknowledge the support of my promoter Michel van Eeten who has patiently guided me through the process of my studies and has meticulously help in improving them.

ABOUT THE AUTHOR

Arman Noroozian was born in Tehran, Iran. He earned his BSc degree in computer science at the University of Tehran, Iran, in 2005 on the subject of bio-inspired and evolutionary software engineering. He later moved to the Netherlands and earned his MSc degree in computer science from Delft University of Technology, Netherlands, in 2010 on the subject of modeling and improving incentives in Peer-to-Peer systems by employing game theoretical concepts and mechanism design. During and after his studies he was employed and developed software for several companies in the Netherlands. He later returned to Delft University of Technology in 2012 as a researcher firstly conducting research in the area of artificial intelligence, and later shifting his research to information security in 2014 when he started his PhD research under the supervision of Prof. Michel van Eeten on empirical security measurement in the hosting market. He has recently become a proud parent and currently works as a post-doctoral researcher in his former research group at the Faculty of Technology, Policy and Management, where he is researching various topics closely related to his PhD research and within the field of information security. A selective list of his co-authored publications may be found hereafter. Arman has a range of research interests including but not limited to empirical studies in security and privacy, as well as applications of economic theories, statistics and AI in networks and in particular around the increasingly important subject of misinformation.

PUBLICATIONS

- A. Noroozian, M. Korczyński, S. TajalizadehKhoob, and M. Van Eeten, "Developing security reputation metrics for hosting providers," in Proceedings of the 8th Workshop on Cyber Security Experimentation and Test (CSET). Berkeley, USA: USENIX Association, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?id=2831120.2831125
- S. Tajalizadehkhoob, M. Korczyński, A. Noroozian, C. Gañán, and M. van Eeten, "Apples, oranges and hosting providers: Heterogeneity and security in the hosting market," in Proceedings of the IEEE/IFIP Network Operations and Management Symposium (NOMS), Istanbul, Turkey: IEEE, 2016, pp. 289–297. [Online]. Available: http://ieeexplore.ieee.org/docum\ent/7502824/
- A. Noroozian A., M. Korczyński , C. Gañán, D. Makita, K. Yoshioka, M. van Eeten, "Who Gets the Boot? Analyzing Victimization by DDoS-as-a-Service". in: Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses (RAID) 2016. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-45719-2_17
- A. Noroozian, M. Ciere, M. Korczyński, S. Tajalizadehkhoob, and M. van Eeten, "Inferring the security performance of providers from noisy and heterogenous abuse datasets," in Proceedings of the 16th Annual Workshop on the Economics of Information Security (WEIS), 2017. [Online]. Available: https://pure.tudelft.nl/portal/files/28435654/WEIS_2017_paper_60.pdf
- M. Korczyński, S. Tajalizadehkhoob, A. Noroozian, M. Wullink, C. Hesselman, and M. v. Eeten, "Reputation metrics design to improve intermediary incentives for security of tlds," in Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P), 2017, pp. 579–594. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7962004/
- S. Tajalizadehkhoob, C. Gañán, A. Noroozian, and M. v. Eeten, "The Role of Hosting Providers in Fighting Command and Control Infrastructure of Financial Malware," in Proceedings of the ACM Asia Conference on Computer and Communications Security (ASIA CCS). Abu Dhabi, UAE: ACM, 2017, pp. 575–586. [Online]. Available: http://doi.acm.org/10.1145/3052973.3053023

- S. Tajalizadehkhoob, T. van Goethem, M. Korczyński, A. Noroozian, R. Böhme, T. Moore, W. Joosen, and M. van Eeten, "Herding Vulnerable Cats: A Statistical Approach to Disentangle Joint Responsibility for Web Security in Shared Hosting," in Proceedings of the ACM Conference on Computer and Communications Security (CCS). ACM, 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3133956.3133971

- M. Korczyński, M. Wullink, S. Tajalizadehkhoob, G. C. M. Moura, A. Noroozian, D. Bagley, and C. Hesselman. "Cybercrime After the Sunrise: A Statistical Analysis of DNS Abuse in New gTLDs". In Proceedings of the 2018 on Asia Conference on Computer and Communications Security (ASIACCS '18). Association for Computing Machinery, New York, NY, USA, 609–623. [Online]. Available: https://doi.org/10.1145/3196494.3196548

- A. Noroozian, J. Koenders, E. Van Veldhuizen, C. Gañán, S. Alrwais, D. McCoy, and M. Van Eeten. "Platforms in everything: analyzing ground-truth data on the anatomy and economics of bullet-proof hosting". In Proceedings of the 28th USENIX Conference on Security Symposium (SEC'19). 2019. USENIX Association, USA, 1341–1356. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/noroozian

- R. van Wegberg, F. Miedema, U. Akyazi, A. Noroozian, B. Klievink and M. van Eeten. "Go see a specialist? Predicting cybercrime sales on online anonymous markets from vendor and product characteristics". In Proceedings of the International Conference on World Wide Web (WWW '20), 2020. [To Appear].