Rotten Apples or Bad Harvest? What We Are Measuring When We Are Measuring Abuse

SAMANEH TAJALIZADEHKHOOB, Delft University of Technology
RAINER BÖHME, University of Innsbruck
CARLOS GAÑÁN, MACIEJ KORCZYŃSKI, and MICHEL VAN EETEN,
Delft University of Technology

Internet security and technology policy research regularly uses technical indicators of abuse to identify culprits and to tailor mitigation strategies. As a major obstacle, current inferences from abuse data that aim to characterize providers with poor security practices often use a *naive* normalization of abuse (abuse counts divided by network size) and do not take into account other inherent or structural properties of providers. Even the size estimates are subject to measurement errors relating to attribution, aggregation, and various sources of heterogeneity. More precise indicators are costly to measure at Internet scale. We address these issues for the case of hosting providers with a statistical model of the abuse data generation process, using phishing sites in hosting networks as a case study. We decompose error sources and then estimate key parameters of the model, controlling for heterogeneity in size and business model. We find that 84% of the variation in abuse counts across 45,358 hosting providers can be explained with structural factors alone. Informed by the fitted model, we systematically select and enrich a subset of 105 homogeneous "statistical twins" with additional explanatory variables, unreasonable to collect for *all* hosting providers. We find that abuse is positively associated with the popularity of websites hosted and with the prevalence of popular content management systems. Moreover, hosting providers who charge higher prices (after controlling for level differences between countries) witness less abuse. These structural factors together explain a further 77% of the remaining variation. This calls into question premature inferences from raw abuse indicators about the security efforts of actors, and suggests the adoption of similar analysis frameworks in all domains where network measurement aims at informing technology policy.

CCS Concepts: • **Security and privacy** → **Economics of security and privacy**; **Systems security**; • **General and reference** → *Metrics*;

Additional Key Words and Phrases: Statistical modeling, hosting providers, abuse concentrations, web security, measurement errors

This work was supported by NWO (Grant No. 12.003/628.001.003), the National Cyber Security Center (NCSC) and SIDN, the .NL Registry, and Archimedes Privatstiftung, Innsbruck.

Authors' addresses: S. Tajalizadehkhoob, C. Gañán, and M. van Eeten, Department of Multi-actor systems, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, the Netherlands; emails: {S.T.Tajalizadehkhoob, c.hernandezganan, m.j.g.vaneeten}@tudelft.nl; R. Böhme, Security and Privacy Lab, University of Innsbruck, Technikerstraße 21A, 6020 Innsbruck, Austria; email: rainer.boehme@uibk.ac.at; M. Korczyński, LIG Lab, Grenoble Institute of Technology, 38401 Saint-Martin d'Heres, France; email: maciej.korczynski@univ-grenoble-alpes.fr.

ACM 1533-5399/2018/07-ART49 $15.00
https://doi.org/10.1145/3122985

ACM Reference format:
Samaneh Tajalizadehkhoob, Rainer Böhme, Carlos Gañán, Maciej Korczyński, and Michel Van Eeten. 2018. Rotten Apples or Bad Harvest? What We Are Measuring When We Are Measuring Abuse. *ACM Trans. Internet Technol.* 18, 4, Article 49 (July 2018), 25 pages.
https://doi.org/10.1145/3122985

1 INTRODUCTION

Abuse data is an important foundation for security and policy research. It associates technical identifiers—typically IP addresses, domain names or URLs—with malicious activities, such as spam, infected machines, command-and-control servers, and phishing sites.

Scientific studies and industry reports draw on abuse data to make inferences about the security practices of the parties in charge of the networks or services where the abuse is located. Concentrations of abuse are seen as evidence of poor security practices or even criminal behavior, explicitly or implicitly characterizing certain providers as "rotten apples" or at least as actors who can and should do more remediation (HostExploit 2017; Goncharov 2015; Hao et al. 2013; Czyz et al. 2014; Stone-Gross et al. 2009; Shue et al. 2012; Noroozian et al. 2015).

Industry representatives often counter these kinds of incomplete/wrong inferences based on right abuse data, when they make media headlines. For example, a 2013 McAfee report ranked the Netherlands as number three worldwide in terms of hosting botnet command-and-control (C&C) servers (McAfee Intel Security 2013). A leading news organization concluded: "Netherlands Paradise for Cybercriminals," prompting a debate that reached the national parliament (Nederlandse Omroep Stichting 2013). The Dutch Hosting Provider Association responded that it "disagrees vehemently" with this conclusion (Dutch Hosting Provider Association 2013). It argued that the high ranking for hosting C&C servers was an artefact of the large hosting infrastructure in the country, not of any negligence or malice on the part of providers.

The hosting provider association raised a valid point. We know that concentrations of abuse are, to a large extent, a function of the size of the network and the service portfolio of the provider, rather than being indicators of the provider's security effort (Clayton et al. 2015).

Previous research looked into the effect of size—as one of the providers' structural properties—on abuse levels (Stone-Gross et al. 2009; Shue et al. 2012). A common problem in these studies is that size is controlled for by dividing the number of abuse events by the number of IP addresses associated with an Autonomous System (AS). This is not only a naive normalization approach, considering all other size indicators that could influence abuse counts, but also contains errors in aggregation and attribution of abuse. It is also problematic to use Autonomous Systems (ASes), which are entities responsible for routing IP addresses, as proxies for hosting providers, the organizations who operate the IP space. Moreover, previous work does not take into account other inherent or structural properties of providers, such as pricing strategies or the type of the hosting service offered. All of these can potentially influence abuse (Liu et al. 2015).

Therefore, to advance our ability of making reliable inferences from abuse data and to address the limitations of prior work, in this article, we develop an analytical approach and propose a statistical model of the abuse data generation process. The model helps to understand to what extent abuse levels are determined by structural properties of providers versus being mainly determined by other factors, including, but not limited to, the security efforts of individual providers. Structural properties in our study are different size and business model variables, pricing strategy, time-in-business, popularity index, WordPress use, and information and communication technology (ICT) development index.

We use phishing abuse data as a case-study to demonstrate our approach. In short, this article makes the following contributions:

- We present a scalable approach to reduce attribution error in studying abuse across the population of hosting providers using passive DNS and WHOIS data;
- We propose an analytical and statistical approach to explain abuse concentrations. Our model improves on previously utilized naive normalization methods. This is done through decomposing the different sources of variance present in abuse data, such as providers' characteristics, attacker behaviour, and measurement errors;
- In a case study on phishing data, we show that more than 84% of the variance in abuse data can be explained by four size and business model properties of providers, collected for the entire population of hosting providers;
- We develop an approach to measure the impact of factors that are difficult to observe at scale. Using "statistical twins," we present the first empirical evidence of the impact of pricing, time-in-business on phishing abuse. Together with other factors related to providers' business models, we are able to explain a further 77% of the remaining variation in abuse, while controlling for country-level differences;
- We demonstrate how our approach generates comparative abuse metrics by controlling for the structural differences among providers. Such metrics are more suitable to evaluate policy impact on concentrations of abuse than absolute counts or naively normalized metrics;

While our study provides an unprecedented view into the interpretation and attribution of abuse in the case of hosting providers, a limiting factor is that we measure only structural properties. We obtain security effort, the variable of key interest, as an unobserved residual, which is conflated with measurement noise. As a result, this indirect approach leaves us with upper-bound estimates of the effect of security effort on abuse concentrations.

Section 2 revisits related work, structured by the level of analysis. In Section 3, we outline the analytical approach that sets up the rest of the paper. Section 4 describes data sources and collection methods. Section 5 details our general modeling approach and present results for the entire sample of hosting providers. In Section 6, we explain the "statistical twins" approach and present results for the subset of enriched data points. In Section 7, we evaluate the robustness by assessing the impact of measurements errors in the abuse data and size estimates. Finally, we discusses our main conclusions and implications in Section 8.

2 RELATED WORK

There is a large body of work on methods for detecting abused resources on the Internet. Observational data on abuse incidents is the starting point for our study, but we do not engage with the detection methods themselves and therefore will not survey them here.

Blacklists as source of observational data on abused resources have already been extensively studied (Pitsillidis et al. 2012; Kührer et al. 2014; Metcalf and Spring 2013). Closer to our research is the line of work that identifies and explains patterns in abuse data. The patterns are studied at different levels of analysis: (i) individual resource (host, IP address, domain); (ii) network, or other aggregates of individual resources; (iii) actor, meaning the economic entity providing the resources or otherwise responsible for them; and (iv) country. We briefly survey relevant work at each level.

Individual resources. A variety of studies have been successful at explaining or predicting the occurrence of abuse, such as compromised websites, from properties of individual resources, such as content-specific features, webserver software versions, or website misconfigurations (Soska and Christin 2014; Liu et al. 2015; Vasek et al. 2016). The factors identified in these studies impact the distribution of abuse on domain names, e.g., domains with outdated software are abused more than others. Others used DNS characteristics to predict whether domain names are malicious (Bilge et al. 2011; Hao et al. 2011). More tailored toward phishing, some authors propose to detect phishing websites using URL and content-specific features (Rosiello et al. 2007; Whittaker et al. 2010).

Networks. Another strand of work looks at how abuse is distributed across aggregate units of resources, such as address blocks, Autonomous Systems (ASes) (Levchenko et al. 2011; Noroozian et al. 2015), or top-level domains (TLDs) (Aaron and Rasmussen 2015a). These studies often identify concentrations of abuse in certain networks (Ramachandran and Feamster 2006; Collins et al. 2007; Levchenko et al. 2011) and then identify network features that correlate with abuse rates, such as poor network hygiene (Zhang et al. 2014; Stone-Gross et al. 2009) or rapidly changing network connectivity (Shue et al. 2012; Wagner et al. 2013; Konte et al. 2015). These studies aim to detect malicious or poorly managed networks, rather than disentangling the factors explaining the causal relationship of why abuse is concentrated or how it is distributed across all networks. Furthermore, to be useful for policy and interventions, the aggregated resources need to be attributed to the relevant economic actors rather than a technical entity. For drawing inferences on providers, an explicit attribution method is needed. This takes us to the third level of analysis.

Actors. Actors are the economic entities that operate resources or are otherwise responsible for them. Work at this level has to bridge the gap between the technical identifiers in abuse data, like ASes, and the organizations responsible for specific resources. This is not straightforward as many Internet Service Providers (ISPs), for example, operate multiple ASes (Asghari et al. 2015b), and many hosting providers share an AS with other providers (Tajalizadehkhoob et al. 2016).

Our work is situated at this level. We are not aware of any work that explains abuse patterns across hosting providers. Liu et al. have studied the extent to which organizations' properties, such as symptoms of mismanagement, size of allocated IP space, and corresponding abuse counts can predict data breach incidents. This work is amongst the first studies that predicted incident rates from one structural property of the organizations (i.e., the number of IP addresses allocated) and several effort-related indicators (i.e., indicators of network mismanagement and misconfiguration). However, it does not distinguish between organizations that offer hosting services and other organizations. It also does not model the structural properties of organizations comprehensively (Liu et al. 2015). Somewhat related is a study on security practices of a small sample of hosting providers (Canali et al. 2013). More mature work exists for domain names. Studies have identified which registrar or registries are associated with malicious domains and quantified the effect of different interventions on abuse rates (Hao et al. 2013; Liu et al. 2011).

Our study extends the prior work in a variety of ways. Unlike the existing work, we are not trying to find technical features to correlate with abuse, but we are trying to understand to what extent abuse levels are a function of structural properties of the industry versus being determined by the security efforts of individual providers. For that, first we adopt a better attribution method for identifying the relevant actors, by moving from ASes toward hosting providers, to whom the IP space is allocated. We fit a multivariate statistical model and include a set of important explanatory factors, such as size of IP space, which have been explored before (Liu et al. 2015). Other properties of hosting providers, such as size of domain name space, size of IP space used for web hosting, portion of shared hosting business, hosting price, time in business, are studied for the first time here.

Countries. The highest level of analysis is countries. Work in this area studies the relationship between country-level factors, such as GDP, rule of law and ICT development, and the distribution of abuse, most notably infected hosts (Asghari et al. 2015a; Garg and Camp 2013; Kleiner et al. 2013). In contrast to this research, we take providers as the unit of analysis, because that is where agency is located in terms of fighting abuse. That being said, country-level factors describe institutional differences in the environments of providers that are also relevant to take into account. In our study, we estimate the impact of the ICT development of a country on abuse, while controlling for other unobserved country-level differences using fixed effects models.

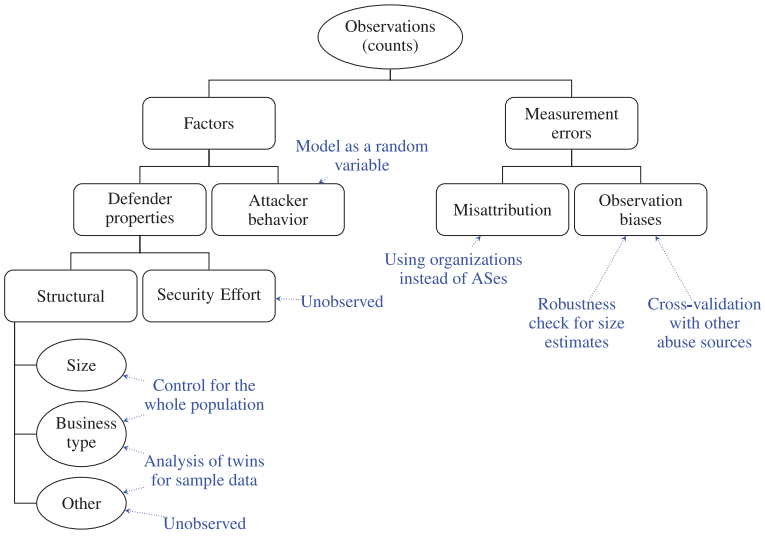


Fig. 1. Overview of our analytical approach.

3 ANALYTICAL APPROACH

To explain the driving forces behind abuse observations at the level of hosting providers, we need to disentangle several sources of variance. Figure 1 summarizes these sources of variance and provides an overview of our approach.

3.1 Decomposition of Sources of Variance

We assume that abuse observations are broadly driven by two phenomena: explanatory factors (left branch) and measurement errors (right branch).

We further divide the explanatory factors into defender properties and attacker behavior. Defender properties are then branched into two main groups: structural properties and (indicators of) security effort. The most relevant structural properties are size metrics. Another structural property is the type of business. Providers offer different types of hosting services, e.g., unmanaged, dedicated, shared, virtual private server, and so on. These services differ in the amount of responsibility the provider assumes and in its role of providing security (M3AAWG 2015). Business type also includes the pricing strategy. Other characteristics include the legal framework under which providers operate as well as the overall maturity of ICT development in a country. Last but not least, providers' security efforts influence the abuse counts.

Next to explanatory factors, we distinguish measurement errors that cause variance in the abuse counts: biases in observations (for factors and abuse data) and problems in attributing abuse incidents to the responsible economic entity.

To asses the influence of each of the above mentioned factors in explaining the variance in the abuse counts of hosting providers, we develop a statistical modeling approach and implement it for one type of abuse data as a case-study: phishing domains. The blue text in Figure 1 indicates how we deal with the various sources of variance as decomposed in the analytical approach.

3.2 Model of the Abuse Data Generating Process

Abuse counts consists of non-negative integers that arise from a direct observation of a point process in a given time period, suggesting to model the data generation process as Poisson distribution. The number of phishing domains per provider enters a Poisson Generalized Linear Model

(GLM) as a dependent variable. Structural defender properties related to size and business type (detailed in Section 5) are used as independent variables. We improve on the size estimates used in previous work in two ways: (i) we improve attribution by calculating size estimates for hosting providers instead of technical entities, like ASes, (ii) in addition to the size of the allocated IP space, we also include the sizes of the domain name space and the IP space used for webhosting in the regression. This allows for more precise control of providers' customer base and attack surface. All these variables can be economically collected for all hosting providers.

Other structural properties (detailed in Section 6) are more cumbersome to collect, since they require manual work or are costly to others when measured at scale. A statistician's response would be to estimate from a random sample. The size of a random sample depends on the target level of confidence and on the effect size (akin the signal-to-noise ratio). In the case of hosting providers, the heterogeneity in the population may hide subtle effects of security effort, which would require uneconomically large samples to control for.

A more efficient approach is to modify the sampling strategy and select subsets of cases that appear homogeneous according to the observable structural properties. Specifically, we select a set of "statistical twins"—subsets of size two—covering the domain of the known population. We collect additional variables for each twin. The subsequent analysis looks for factors explaining differences *within* twins, disregarding differences *between* twins. Technically, this can be achieved by adding one fixed effect per subset to the GLM specification. This method allows us to control for large parts of the heterogeneity and at the same time account for linear bias introduced by the systematic sample selection. Another way of looking at this approach is that we select the a priori most informative cases from the population for further analysis. It rests on the implied assumption that cases that have a twin in the population do not systematically differ from those that do not. We conjecture that this is not unreasonable for the population of global hosting providers.

Of course, there always remain unobserved factors, including additional structural properties, variables related to security effort, attacker behavior, and all kinds of measurement error. Although all unobserved factors are conflated in the unexplained variance of the model, below we explain how we deal with each of these factors.

3.3 Security Effort as Residual

We model security effort as an unobserved variable, because direct measurement of security effort for hosting providers is very difficult. First, there is no way to directly observe the actions of providers' security staff, such as applying security patches, educating customers, and putting application firewalls in place. At best, we can observe some technical indicators that might be influenced by those efforts, but that will always constitute a very partial measurement. Second, even if certain technical indicators can be collected as proxies for effort, they are not necessarily causally related to abuse. Indicators used in prior work measure network hygiene (e.g., BGP misconfiguration, untrusted certificates, open mail relays). This type of hygiene is not preventative of web compromise. At best, such hygiene indicators can be interpreted as measuring some more generic security effort, which might or might not correlate with providers' effort in fighting abuse. Third, useful technical proxies for effort are still hard to attribute to providers. Who actually caused the presence or absence of the vulnerability? The web master, the hosting provider, the software vendor, or someone else? Fourth, and last, even if useful indicators can be observed and correctly attributed, it is hard to draw the correct inference from them because of the heterogeneity in the market. Hosting providers are very different and so are domains within providers. An indicator that signals lack of security effort for one domain and provider might not signal the same for another domain or provider, because the affected users may face very different threats. For instance, not all websites are more secure by adopting HTTP Strict Transport Security (HSTS) or Content

Table 1. Descriptive Statistics of Variables Used in the Full Model for All Providers and Providers in the Sample of Twins

	Min	Mean	Median	Max	SD
For all data points (n = 45,358)					
Number of assigned IPs [log10]	0	3.1	3.2	8.4	1.2
Number of IPs hosting domains [log10]	0	1.8	1.7	6.2	0.8
Number of hosted domains [log10]	0	2.0	1.8	7.6	0.9
Percentage of domains hosted on shared IPs	0	51.0	59.0	100	37.1
Number of phishing domains in APWG	0	2.8	0	11,455	91.3
Number of phishing domains in Cyscon	0	0.9	0	5,515	37.4
For statistical twins (n = 210)					
Number of assigned IPs [log10]	0.3	4.0	4.0	7.5	1.4
Number of IPs hosting domains [log10]	0.3	3.0	3.0	5.6	1.2
Number of hosted domains [log10]	1.5	3.9	3.7	7.6	1.2
Percentage of domains hosted on shared IPs	0	78.6	87.9	99.3	22.3
Number of phishing domains in APWG	0	159.2	3	9,805	967.6
Number of phishing domains in Cyscon	0	54.8	1	3,819	375.0

Security Policy (CSP). And many providers cannot cause BGP misconfigurations, as several they do not operate their own AS and, hence, BGP announcements.

In short, measuring effort is very difficult under the best of circumstances. This makes it valuable to develop an alternative approach to approximate effort that does not depend on direct observation. In this work, we explore the option of treating security effort as residual, i.e., as part of the unexplained variance, after accounting for the observable factors. As a consequence, our results must be interpreted as upper bounds for the effect of security effort on abuse. This introduces some asymmetry in our research logic: only if the residual variance is small, we can rule out that security effort is very influential.

We model attacker behavior as random variable, assuming that most attackers behave rather opportunistic than strategic. The realization of this random variable is also part of the unexplained variance. It remains at the discretion of the reader whether the realization of discrete incidents shall be interpreted as part of the measurement error of an underlying latent variable of "attack strength" or as the result of seemingly random attacker behavior. To manage expectations, we note that our statistical approach is generally not suitable for studying targeted attacks and other rare events.

Avoiding observational biases is outside the scope of this article, but we try to limit their effect on our core results by cross-validation against a different set of phishing data. In addition, we test the robustness of the size estimates in the model against errors due to possible model mismatch by simulating the impact of distorted size estimates against a simulated phishing count variable drawn from an ideal Poisson distribution model.

4 DATA

Our study uses a variety of data sources, which are summarized in Table 1.

4.1 Mapping to Hosting Providers

Our goal is to accurately identify the IP ranges that belong to hosting providers. Most of the existing work uses BGP data to map IP addresses of abuse incidents to ASes, equating the latter with hosting providers. However, the entity that is routing an IP address is not always the same as the

organization that is hosting an IP address. While some organizations operate under several ASes, other organizations share a single AS. Our prior work finds that on average an AS consists of seven organizations (Tajalizadehkhoob et al. 2016). WHOIS registration and IP allocation information, which is collected and stored by Regional Internet Registries (RIR), provide more direct visibility into the responsible organization behind an IP address. It should be noted that WHOIS data comes with its well-known limitations, such as different data formats, they are less detrimental on analysis of the hosting market than starting with BGP (Elliott 2008).

We refined the general method describes by Cai et al. and Liu et al. for mapping IP addresses to their organizations using WHOIS, and extracted only hosting providers from the set of organizations (Cai et al. 2010; Liu et al. 2015). In a nutshell, we only map organizations that host domains seen in DNSDB passive DNS database—a database that draws upon hundreds of sensors worldwide—generously provided to us by Farsight Security (Farsight Security 2016; DNSDB 2016). The DNSDB passive DNS data contains second-level domain names and the IP addresses that they resolved to. To our knowledge, DNSDB has the best coverage of the domain name space available to researchers. We take 47,446,082 IP addresses and 214,138,467 domain names observed in passive DNS data in 2015 and map them to 161,891 corresponding organizations to whom they are allocated, using the MaxMind API for WHOIS IP allocation, as discussed in previous research (MaxMind 2016; Liu et al. 2015; Dimitropoulos et al. 2006). The resulting list contains all organizations to whom IP ranges are allocated. Many of them are not hosting providers, e.g., `Massachusetts Institute of Technology` and `DoD Network Information Center`. We compile a final set of organizations that offer hosting services by filtering out non-hosting providers through a series of keywords related to educational and government-related organizations, ISPs, broadband providers, mobile service providers, domain parking services, and DDoS protection services (Tajalizadehkhoob et al. 2016; Dimitropoulos et al. 2006). The final set consists of 45,358 hosting providers.

4.2 Abuse Data

To demonstrate the application of our proposed analytical method, we model the count of abuse in the networks of hosting providers, using phishing data as a case-study. The main reason behind this choice is, since phishing sites are known to be mostly compromised accounts (Aaron and Rasmussen 2015a), bypassing security is very much required in the bulk of cases. To that end, we analyze phishing domains collected from two sources: the Anti-Phishing Working Group (APWG) (Anti-Phishing Working Group 2016) and Cyscon GmbH (GMBH. 2016).

APWG data contains URLs used in phishing attacks together with their blacklisting times. We collect the IP addresses associated with second-level domains[1] in the APWG feed by retrieving the corresponding passive DNS entry at the time when the domain is blacklisted. The final set consists of 131,018 unique second-level domains and 95,294 unique IP addresses hosted by 5,391 hosting providers for the entire 2015.

Cyscon phishing data contains the same attributes (URLs, IP addresses, blacklisting times). We collect 40,292 unique second-level domains and 23,021 unique IP addresses hosted by 2,782 hosting providers in June–December 2015.

We use the phishing second-level domains in APWG data as the default response variable in our regression models in Sections 5 and 6. In Section 7, we use the phishing second-level domains in Cyscon data to cross-validate the results.

[1]Domains such as `example.co.uk` are considered to be second-level domains as well.

4.3 Provider Properties

To explain the differences in phishing incident counts between hosting providers, we collect a number of variables on provider (defender) properties, some for the entire population and some for the sample of "statistical twins."

4.3.1 Variables Collected for All Providers. In addition to identifying providers, we can collect variables related to size and business model (see the leftmost factors in Figure 1) from passive DNS and WHOIS data.

Number of assigned IP addresses. Size of IP address block(s) assigned to a provider based on WHOIS.

Number of IP addresses hosting domains. Count of IP addresses seen to host domains in passive DNS. The combination of these two variables provides information about the kind of business the hosting provider is running. For instance, providers who use a large part of their assigned IP space for hosting domains such as webhosting providers can have a different business model from providers who use their IP space for hosting other services such as data centers.

Number of hosted domains. Count of the second-level domains in the passive DNS data. In addition, note that since the first three variables have a skewed distribution, we log-transform them with base 10.

Percentage of domains hosted on shared IPs. We consider an IP address shared, if it hosts more than 10 domain names (Tajalizadehkhoob et al. 2016).

This variable measures the ratio of domains that are hosted on shared IP addresses over the total size of the hosted domains, in percent. This variable not only conveys information about the size of the shared hosting infrastructure of the provider but also about the provider's business model: the degree to which it relies on low-cost shared hosting services.

4.3.2 "Statistical Twins" Sampling Method. It is not possible or desirable to collect data at scale for all factors in Figure 1. For example, pricing information must be collected manually. It involves search, interpretation, and human judgment. Applying a standardized procedure is too costly for the entire population of 45,358 hosting providers for some of these variables. Even collecting some technical indicators, such as the number of Wordpress installation on *all* websites of *every* hosting provider, are inefficient to collect in bulk; perhaps even unethical, because it imposes a cost on the scanned networks.

For this reason, we employ a data-driven sampling strategy and select a small set of homogeneous "statistical twins," for which we can collect as much information as possible. The steps are:

(i) We start with a set of randomly and uniformly selected *seed data points* (105 hosting providers), for which we have collected pricing information. Let $\mathcal{S}$ be the set of seed data points and $\mathcal{T}$ be the total set (or population) of providers. The random seed should ensure a good coverage of the population.
(ii) We calculate the distance between all the data points in $\mathcal{S}$ and data points in $\mathcal{T}$ using the Euclidean distance between all explanatory variables collected for the entire population. This results in a distance matrix of 105 rows and around 45 k columns.
(iii) For each of the 105 providers in $\mathcal{S}$, we select the closest match; that is the provider in set $\mathcal{T}$ that has the minimum distance to the provider in set $\mathcal{S}$, in terms of variables in Table 1.
(iv) This results in a set of *rich data points* $\mathcal{R}$ consisting of 105 homogeneous statistical twins and 206 unique hosting providers in total, where a few providers became part of two twins.

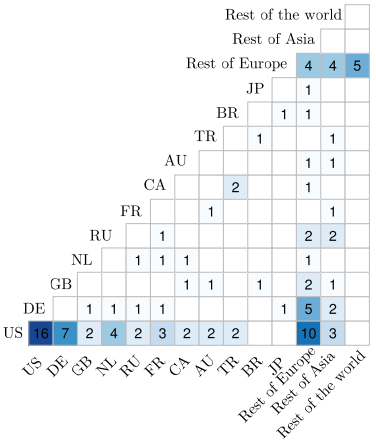


Fig. 2. Number of twins per country combination.

We further use set $\mathcal{R}$ as our (notoriously biased) sample to study the effect of additional factors on abuse. We account for the bias in the analysis (Section 7).

(v) We collect additional variables for all elements in $\mathcal{R}$.

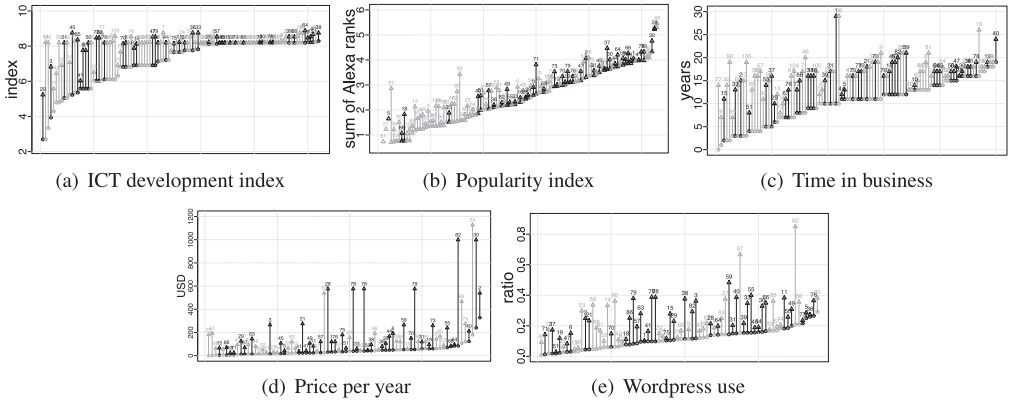
While the method is economical and increases the information gained per effort, it comes with some limitations. First, it is unlikely to spot outliers, simply because there is nothing to pair up against a unique provider like Amazon. Second, it assumes that data points in dense regions of the population do not systematically differ from those in sparser regions, where the probability of finding twins is lower. Third, it requires that the bias introduced by the selection strategy is approximately linear. Non-linear bias correction is possible in principle, but requires prior information on the functional form.

4.3.3 Variables Collected for a Sample of Providers. The below variables are collected only for the providers in $\mathcal{R}$, as explained in Section 4.3.2.

Country. For the providers in $\mathcal{R}$, we use MaxMind API to identify where the majority of IP addresses are located. In Figure 2, we show the general coverage of twins in our data. For instance, there are 16 twins with both providers located in the U.S. while there is 1 twin with one provider located in Britain and the other in Japan. The figure indicates a good variation in the geo-location of twins in our sample data.

ICT development index. The ITU publishes country-level variables that proxy the development in information and communication technology (ICT) using several indicators (International Telecommunication Union (ITU) 2014). (This and the previous variable came into view for our enriched data points, but they could also be economically collected for the whole population.)

Popularity index. We use Alexa's one million top-ranked domains as a proxy for online popularity. A provider is assumed to host more popular content when more domains are on the list. If attacker behavior is not completely random, then one may expect that providers with more popular sites are targeted more to compromise domains and set up phishing sites. The popularity index per provider in calculated by aggregating the Alexa ranks of second-level domains in our sample $\mathcal{R}$. We first reverse the Alexa ranks per domain, that is the most popular Alexa domain gets the rank 1,000,000. We then calculate a score per provider by summing up the base-10 logarithm of the reverse rank of all ranked websites. This score combines website (i.e., customer) popularity and the density of popular sites at a hosting provider and the method allows us to account for extreme popularity of large providers.



(a) ICT development index (b) Popularity index (c) Time in business

(d) Price per year (e) Wordpress use

Fig. 3. Visualization of the variation of additional variables within and between sampled twins (x axis is sorted based on providers with lower value to the ones with higher value in each twin).

Time in business. Time in business is a proxy for how experienced a provider is. We collect this information by querying the WHOIS database for the registration date of the provider's own domain name. Missing values were entered if we could not find the website or there was no public registration date in WHOIS. We cross-checked the results with the Internet Archive for all data points (Web-Archive 2016). Almost all domains in our sample were captured by Web-archive a couple of months after the day they were registered.

Pricing. Finding comparable pricing information for hosting providers is complicated. We visit the provider's website and collect prices for the least expensive hosting plan on offer, as an indicator to tell "bottom-end" and "top-end" apart. We hypothesize that providers with cheaper hosting plans have fewer resources and more vulnerable customers, so higher phishing counts. We converted all prices to US dollars by taking the 2015 average exchange rate of the local currency to USD. Price information for providers is missing if (i) we could not find the provider's website; (ii) the prices are not available online; and (iii) we do not receive a reply to our inquiries through other channels.

WordPress use. Previous studies suggest that domains with popular content management systems in general have higher odds of being compromised for phishing attacks (Vasek et al. 2016). Therefore, we use WordPress as a proxy of popular and targeted software and hypothesize that providers with such business models have higher chance of incurring a phishing incident. More specifically, the more WordPress websites a provider hosts, the higher the chance of a compromise that is used for a phishing site. Note that we collect this indicator to capture information about the business model. One could also collect indicators of software installations and patch levels to measure to measure the impact of provider's patching practices on abuse, but this is outside the scope of this article. To collect data on WordPress installations per provider, we randomly sample 2% of providers' domains in $\mathcal{R}$. This results in 1,398,928 domains. We use WPScan, a black box WordPress vulnerability scanner developed by Sucuri, to determine if a domain uses WordPress (WPScan 2016). The variable "WordPress use" is calculated by taking the ratio of domains scanned with a WordPress installation over all domains scanned, excluding those that we were unable to scan.

To raise confidence in the selection of twins and facilitate the interpretation of estimated coefficients, in Figure 3 we show a collection of plots visualizing the variance within and between twins for the additional variables. Circles display value of the provider with lower value in a twin

and are the basis for sorting all twins horizontally. Triangles display the respective value of the provider with higher value.

In general, the plots suggest a broad enough coverage of data points in the twins. More specifically, as Figure 3(a) demonstrates, variance of the ICT development index decreases with increasing the base level (The provider shown with circle, those with lower value in a twin). This is probably due the concentration of hosting providers (and hence twins) in a few large and highly developed countries, as witnessed in the country pairings displayed in Figure 2. The popularity index exhibits moderate differences at all levels (Figure 3(b)). This indicates that the sampling strategy accounts well for the heterogeneity in the size of hosting providers, which is also included in the index calculation. At the same time, the remaining variation allows for the statistical identification of potentially influencing factors within the twins. Also, time in business is very "healthy" in this regard, with a smaller difference for values between providers in twins that are in business since the `.com` ages (Figure 3(c)). The differences in price are rather small and exhibit occasional spikes (Figure 3(d)). This may reflect the generally low cost of the cheapest package of one of the twins, differences in business models at the spikes, and potential issues related to comparing U.S. dollar amounts among countries with very different labor and infrastructure costs. Finally, the Wordress use also shows a good mix of variation within and between twins, increasing the chance of extracting a meaningful signal if the indicator has explanatory power (Figure 3(e)).

The general coverage of the rich data points $\mathcal{R}$ of the population $\mathcal{T}$ is best assessed by comparing the descriptive statistics of the four explanatory variables and the response variables, which are available for the entire population. This information is included in the lower half of Table 1.

5 MODELING PHISHING COUNTS

We now propose a statistical model to analyze the extent to which structural properties of hosting providers can explain the concentration of abuse in their networks, for the case of phishing abuse.

5.1 Regression Model

Abuse is measured by counting abuse events per provider. These counts consists of non-negative integers that arise from a direct observation of a point process. In a minimal model, the underlying process is assumed to be stationary and homogeneous, with i.i.d. arrival times for abuse events and thus can be modeled with a Poisson distribution. In Section 5.2, we explain in more details the reason why we opted for Poisson regression over other Poisson-family models, such as Quasi-Poisson and Negative Binomial.

Let us define our response variable Y_i as the number of *abused* second-level domains hosted by provider i, for $i = 1, \ldots, n$, with n being the total number of hosting providers. Let Y_i follow a Poisson distribution with parameter $\lambda \geq 0$, The Poisson distribution has equal mean and variance $E[Y_i] = var[Y_i] = \lambda_i$. In the log-linear version of the GLM, λ_i is modeled as

$$\ln(\lambda_i) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j, \tag{1}$$

where β_0 is the intercept, x_{ij}, $j = 1, \ldots, k$, are explanatory variables representing the structural properties that drive the response variable Y_i, and β_j are parameters to be estimated with maximum likelihood (ML). Statistical hypothesis tests can tell if a parameter β_j significantly differs from zero. If the null hypothesis is rejected, then the corresponding explanatory variable is considered influential and the parameter's sign and magnitude can be interpreted.

5.2 Model Goodness of Fit

The fitted values produced by inserting the ML estimates $\hat{\boldsymbol{\beta}}$ into Equation (1) will not match the values of the phishing data perfectly, chiefly because the data points are realizations and the fitted values are parameters of Poisson distributions. The discrepancy between the model and the data is a measure of the inadequacy of the model. Several measures exist to assess the goodness of fit of GLMs such as Log-likelihood, Akaike Information Criterion (AIC), dispersion parameter of the Poisson model and R-squared. Here we discuss a few of them that are more specific to our Poisson model.

Over-dispersion. Recall that the Poisson model assumes equal mean and variance for the response variable, that is $var[Y_i] = \phi E[Y_i] = \phi\lambda_i$, with $\phi \stackrel{!}{=} 1$, where ϕ is a dispersion parameter. However, this assumption is often "violated" in practice; that is, a likelihood function that leaves ϕ as a parameter to be estimated ($\hat{\phi}$) fits the data much better. In case of heterogeneous count variables, $\hat{\phi} > 1$ indicates signs of over-dispersion, which can be interpreted as unobserved heterogeneity in terms of a missing structural factor that leads to concentrations of observable events.

One might approach over-dispersion by starting from a Poisson model and adding a multiplicative random effect to represent unobserved heterogeneity. This leads to a Negative Binomial GLM. However, even if both parameters of the assumed Negative Binomial distribution are correctly specified, if the distribution of the response variable is not in fact the negative binomial, the maximum-likelihood estimator becomes inconsistent (Cameron and Trivedi 1990). To make sure this holds for our data as well, we have constructed other models that control for over-dispersion, such as Quasi-Poisson and Negative Binomial models, and observed that all significant relationships stayed the same, with minor or no variation in the coefficients and minor variations in standard error values.

The literature suggests that in the absence of a precise mechanism that produces the over-dispersion, it is safe to assume $var(Y_i) = \phi\lambda_i$, for positive values of ϕ. This approach is generally considered robust, since even significant deviations have only a minor effect on the fitted values, their standard errors, confidence intervals, and p-values of hypothesis tests (Heinzl and Mittlböck 2003). Moreover, over-dispersion, is a sign of unobserved heterogeneity and is an indicator for structural variance in our model. Any attempt to compensate it with the choice of more complex distribution functions, such as negative binomial or zero-inflated Poisson, may hide the very signal we are looking for.

The dispersion parameter ϕ of a Poisson regression model is calculated using the chi-square statistics χ^2 divided by degrees of freedom, as it is more robust against outliers (Mittlböck 2002).

$$\hat{\phi} = \frac{\chi^2}{(n-k-1)} = \sum_i \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i \cdot (n-k-1)}, \tag{2}$$

with n being the number of observations and k the number of coefficients. $\boldsymbol{y} = (y_1, \ldots, y_n)'$ are the observed values of the response variable; $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_n)'$ are the corresponding predicted values under the fitted model.

Pseudo R-Squared. A popular measure to assess the fraction of variance explained by a linear model is the R-squared statistic. Similar statistics that take values between 0 (not better than intercept-only model) and 1 (perfect fit) have been proposed for GLMs and are called *pseudo R*-Squared. According to Heinzl and Mittlböck (2003), a pseudo R-Squared measure for Poisson models that takes the effect of over-dispersion into account is given by

$$R^2 = 1 - \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\lambda}}) + k \cdot \hat{\phi}}{D(\boldsymbol{y}, \bar{Y})}, \tag{3}$$

Table 2. GLM for Count of Phishing Domains in APWG for All the Hosting Providers

	Response variable: Count of phishing domains				
	Poisson with log link function				
	(1)	(2)	(3)	(4)	(5)
Number of assigned IP addresses		1.186***	−1.665***	−0.728***	−0.768***
		(0.002)	(0.006)	(0.006)	(0.006)
Number of IP addresses hosting domains			3.623***	1.104***	1.570***
			(0.006)	(0.008)	(0.010)
Number of hosted domains				1.686***	1.238***
				(0.004)	(0.006)
Percentage of domains hosted on shared IP addresses					0.027***
					(0.0003)
Constant	−0.122***	−3.594***	−2.732***	−5.072***	−6.755***
	(0.005)	(0.010)	(0.011)	(0.014)	(0.024)
Observations	45,358	45,358	45,358	45,358	45,358
Log likelihood	−223,113.400	−514,546.600	−236,442.400	−117,601.700	−111,570.800
Akaike Inf. Crit.	446,228.800	1,029,097.000	472,890.800	235,211.400	223,151.700
Dispersion	2934.775	619.708	554.695	12.149	13.166
Pseudo R^2		0.221	0.648	0.831	0.841

Note: Standard errors in brackets.
*p<0.05; **p<0.01; ***p<0.001.

where $D(\boldsymbol{y}, \hat{\boldsymbol{\lambda}})$ is the deviance of the fitted model, $D(\boldsymbol{y}, \bar{Y})$ is the deviance of the intercept-only model, $\hat{\phi}$ is the estimated dispersion (Equation (2)), k is the number of covariates fitted (excluding intercept) and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

5.3 Model Specifications

After selecting the proper regression model and discussing goodness of fit measures, we choose to fit different specifications of the model with a step-wise inclusion of the variables that capture providers' structural properties. For each of the variables, we hypothesize the direction of its relation with phishing counts.

We expect that the number of phishing counts increases as the "Number of hosted domains" and "Number of IP addresses hosting domains" increase. Both variables are correlated and measure some aspects of the size of a provider. We may expect that the coefficient sizes decline if both enter the model together, but it is unlikely that one of them becomes redundant. In the case of "Number of assigned IP addresses," the assumption is slightly different, since the more assigned IP addresses does not necessarily mean that the provider uses them for web hosting. In contrast, we found that the business model of providers with larger assigned IP space is closer to a broadband provider who uses only a very small portion of its assigned space for web-hosting. Accordingly, since phishing attack—as an instance of abuse—directly depends on web-hosting, we expect providers with large "Number of assigned IP addresses," to have less phishing events in a specification where size is already controlled for with the two other variables.

In addition, note that our log-transformation of the top three variables shown in Table 2, perfectly matches with the log-link function of the Poisson model.

We expect that the variable "Percentage of domains hosted on shared IP addresses" correlates positively with the phishing counts of providers due to commonly known vulnerabilities of shared

hosting services (Nikiforakis et al. 2011; Aaron and Rasmussen 2015b), assuming that attackers would go for targets that are easier to compromise.

5.4 Estimation Results

We construct several models by performing a step-wise inclusion of the explanatory variables explained in Section 5.3. A summary of these regression models is presented in Table 2. The table contains five models with estimated coefficients of explanatory variables in each model. Each coefficient demonstrates the amount of variance in the phishing counts that a variable can explain given other variables in the model.

As we move from model (2) to (5), we aim to improve our models in terms of goodness of fit metrics explained in Section 5.2. Likewise, the pseudo R-squared values increase when explanatory variables are added from model (2) to (5). More specifically, with the four size and business-related variables, we are able to explain 84% of the variance in abuse counts across 45,358 hosting providers with simple structural properties of providers. One should note that the dispersion parameter $\hat{\phi}$ across the models has been reduced from 2934.77 for the intercept-only model to 13.17 in model (5). In addition to other factors, the significance and signs of the estimated coefficient for explanatory variables do not change between models (3) and (5), which further indicates that we can safely take model (5) as our final model.

The results in model (5) clearly suggests that the number of phishing sites increases as the "Number of hosted domains" and "Number of IP addresses hosting domains" increases. One should note that these two variables together best capture the *attack surface* of the hosting providers for the case of phishing attacks. Hence, the results demonstrates that for larger attack surfaces, there are more phishing websites.

As hypothesized, "Percentage of domains hosted on shared IP addresses" show a statistically significant relation with the abuse counts, indicating that having more domains on shared servers make a provider more exposed to phishing attacks. The "Number of assigned IP addresses" shows a statistically significant negative relation with the abuse count, as expected when controlling for size. As pointed out earlier, a manual inspection of the providers with large "Number of assigned IP addresses" suggests that they are mostly broadband providers who offer web hosting only as a very small part of their business. Therefore, the negative sign of "Percentage of domains hosted on shared IP addresses" works in line with our hypothesis of having more web hosting domains and IPs as attack surface, increases the number of phishing victims.

In addition, the coefficients and pseudo R-squared values in models (2) to (5) further confirm the point we made earlier in the introduction of the article that a simple bi-variate correlation or a *naive* normalization of abuse with one size metric, while neglecting other size properties, cannot comprehensively explain the variance in abuse counts.

Taking model (2) as an example, the value of estimated coefficient for "Number of assigned IP addresses" suggests that increasing this variable by one unit (i.e., one decimal order of magnitude), multiplies the number of expected abuse counts by $e^{1.186} = 3.273$. However, a *naive* normalization of abuse (abuse counts divided by network size) would have assumed a coefficient equal to 1 for the "Number of assigned IP addresses." Here, our study distances itself significantly from the prior work, where just one size metric is taken into account. In the multivariate models, several size indicators offset each other, making the estimation more precise. In addition, in model (2) where "Number of assigned IP addresses" is the only size variable, we are only able to explain 22% of the variance in phishing counts, whereas by adding three other size metrics, the explained variance is improved up to 84%.

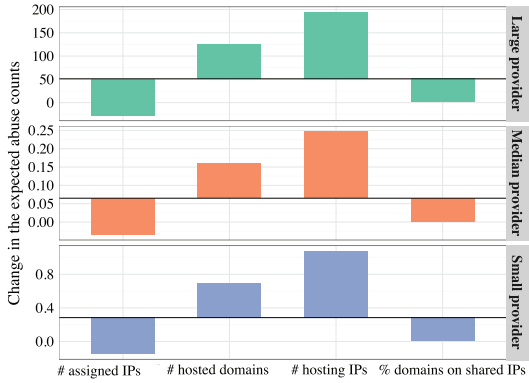


Fig. 4. Partial effect of one unit increase in the predictors on the expected phishing counts in model (5) of Table 2 (note the different scales).

5.5 Quantitative Interpretation

To illustrate the economic significance of the parameters in the fitted model (5), we familiarize the reader with three scenarios. The scenarios are based on hosting provider groups discussed in Tajalizadehkhoob et al. (2016), which is medium, small, and large hosting providers.

For each scenario, we show the partial effect of changes in the abuse counts given each of the explanatory variables (see Figure 4). In the first scenario (scenario 1), all explanatory variables are set at their median as a baseline situation (see Figure 4). In scenario 2, the baseline shows a typical small shared hosting provider with a small number of assigned IP addresses (0.47: $\approx 10^{0.47}$ IPs assigned), a small number of IP addresses used for hosting domains (0.47: $\approx 10^{0.47}$ IPs), a small number of hosted domains (1.95: $\approx 10^{1.95}$ domains), and a high percentage of domains hosted on shared IP addresses (100%). In scenario 3, the baseline situation indicates a large web hosting provider with a huge assigned IP address space (6.85: $\approx 10^{6.85}$), a large IP address space used for hosting domains (5.67: $\approx 10^{5.67}$ IPs assigned), a large amount of hosted domain names (5.68: $\approx 10^{5.68}$ domains), and a very small percentage of domains hosted on shared IP addresses (0.48%). Apparently, this is the common case for web hosting providers that are mostly offering dedicated services (Tajalizadehkhoob et al. 2016).

The bars for each of the scenarios in Figure 4 illustrate the change in the expected count of abuse events for a unit change in each of the explanatory variables, while holding the others constant. Given the coefficients for the explanatory variables in Model (5) in Table 2, one can easily observe that changes in "Number of IP addresses hosting domains" while other variables are constant, changes the phishing counts the most, while the effect of one unit change in "Percentage of domains hosted on shared IP addresses" on the phishing counts is small although statistically significant.

Obviously, evaluating how effective providers are in keeping phishing sites off their networks is important for developing policies and best practices. Absolute concentrations of abuse are still useful, of course. Even if they reflect structural factors, those providers could be asked to go beyond the security practices of the industry, because of their prominent position. Such a call is less plausible, however, if attackers can easily switch among the thousands of providers, as has been found in other work (Levchenko et al. 2011).

Rather than looking at absolute counts of abuse alone, measuring the amount of abuse relative to the providers' structural properties adds valuable information. Figure 5 visualizes the difference

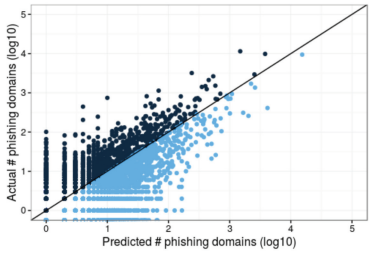


Fig. 5. Observed and predicted number of phishing domains for Model (5) in Table 2.

between actual phishing counts and the counts predicted by our model. Providers below the line are performing better than average and the distance tells us by how much.

6 ADDITIONAL PROVIDER STRUCTURAL PROPERTIES

In model (5) of Table 2, we see that around 84% of the variance in phishing counts is explained by a number of the structural properties of providers, namely, four variables related to size and business model of the hosting providers.

In this section, we continue with interpreting regression models, this time on a sample of the rich data points (set $\mathcal{R}$) as described in Section 4.3.2. Recall that we need to adjust the estimation method by introducing two sets of fixed-effects, (i) for level differences between statistical twins and (ii) for level differences between countries. Fixed effects take away known linear dependence in the residuals. This is essential to obtain accurate test statistics (which assume independent residuals). Sources of dependence are within twins due to the selection strategy and within countries due to the inclusion of country-level variables.

We define the model with (one set of) fixed-effects as

$$\ln(\lambda_{ij}) = \beta x_{ij} + \cdots + \delta_i, \tag{4}$$

where β is the estimated coefficient for x_{ij}, x_{ij} are explanatory variables collected for hosting providers in set $\mathcal{R}$ and δ_i is the "fixed-effect" parameter (Cameron and Trivedi 2013). Subscript i refers to different twins and $j \in \{1, 2\}$ refers to different measurements within each twin, i.e., the same variable measured at different hosting providers belonging to the same twin.

The model uses variables explained in Section 4.3.3 as predictors. We add two fixed-effects to the model—twin and country—by fitting a separate dummy variable as a predictor for each class of each variable. The twin fixed effect controls for the bias introduced by selecting similar samples. The country fixed effect prevents undue dependence in the residuals if country-level predictors are included. In addition, in case of missing values per explanatory variable, we perform a list-wise exclusion on a twin, if one of providers is missing. This further reduces the number of pairs per model depending on missing values of the included variables.

With fixed effects, the definition of pseudo R-squared requires some reflection. It is possible to use an intercept-only baseline, which results in artificially high pseudo R-squared values, because the level differences of the fixed effects are counted as "explained." A more conservative measure is to use the fixed-effects-only model as baseline. Table 3 shows the summary of our results.

Table 3 contains two baseline models (models (1) and (2)). While the former only models twins as a fixed-effect, the later models both twins and countries as fixed effects. Model (3) broadens model (2) by fitting more explanatory variables with fewer missing values. In model (4), we add all the explanatory variables collected for the set of rich datapoints ($\mathcal{R}$), except for the ICT

Table 3. GLM for Count Phishing Domains in APWG for the "Statistical Twins"

	Response Variable: Count of phishing domains				
	Poisson with log link function				
	(1)	(2)	(3)	(4)	(5)
Price per year				0.0003	-0.007^{***}
				(0.0002)	(0.001)
Popularity index (in thousands)			0.001^{***}	0.02^{***}	0.1^{***}
			(0.000)	(0.002)	(0.01)
Time in business			-0.017^{*}	-0.059^{***}	0.015
			(0.007)	(0.005)	(0.012)
ICT dev. index			0.951^{***}		−165.065
			(0.214)		$(>10^3)$
Wordpress use				5.858^{***}	2.203^{***}
				(0.271)	(0.450)
Twin fixed-effects	Yes	Yes	Yes	Yes	Yes
Country fixed-effects	No	Yes	Yes	No	Yes
Observations	210	210	180	84	82
Log likelihood	−2,783.157	−1,111.825	−966.249	−795.838	−249.780
Akaike Inf. Crit.	5,776.315	2,521.650	2,192.499	1,683.677	641.560
Dispersion	40.133	25.770	27.352	31.554	11.243
Pseudo R^2			-0.055	0.625	0.772
Total pseudo R^2	0.974	0.991	0.992	0.966	0.995

Note: Standard errors in brackets.
*p<0.05; **p<0.01; ***p<0.001

development index, having only twins as a fixed effect. In addition to variables in model (4), in model (5) we add ICT development index, setting both twin and country as fixed effects.

The regression results in Table 3 indicate that by including both twins and country as fixed effects, keeping in mind that the datapoints in set $\mathcal{R}$ are already homogeneous in terms of other variables, we are able to explain around 77% of the variance in phishing counts for the twins (pseudo R-squared value). Note that this variance is the 77% of the 16% unexplained variance that is remained in model (5) of Table 2, for the full population of providers. The results further highlight the importance of accounting for other—non-size related—structural properties of providers, while explaining the variance in concentration of abuse.

Even though we have around 100 fixed-effects, we still get very clear and significant results for the main effects. In addition, by moving from model (1) to model (5) in Table 3, we are reducing the amount of unexplained heterogeneity (model's dispersion) from 40.133 to 11.243.

As hypothesized, we see a significant negative relation between the price of hosting and phishing counts in model (5). That is, if we increase price by one unit holding the other variables constant, the phishing count will be multiplied by $e^{(-0.007)} = 0.99$, which means that the cheaper a service is, the more the hosting provider is prone to phishing attacks. Interestingly, variable "Price per year" shows a different relation in model (4), where we do not control for cross-country differences. This change is expected, however, as properties of hosting markets in different countries can differ a lot, which eventually influences the cost structure of the country in regards to hosting services. In addition, the cost of a hosting plan is relative to the economy of the provider's country. Conversion of prices in a specific country to USD, if not controlled for the country differences,

can be very crude. Variables "Wordpress use" and "Popularity index" also show a significant positive relation with phishing count indicating that more Wordpress sites per provider, as a proxy for providers that host oft-attacked software, translates to more phishing attacks, which is quite logical considering the fact that the majority of phishing sites are on compromised machines. One unit increase in "Wordpress use" while holding other variables constant, multiplies the phishing counts by $e^{(2.203)} = 9.052$. Similarly, the more popular websites a provider hosts (popularity index), the more that provider is a victim of phishing attacks.

For "ICT development index," we observe both a sign and significance change from model (3) to (5). This can be understood by looking back at the distribution of twins in Figure 3(b), where the gray color marks the twins that were excluded from model (5) due to missing values. From the figure it is visible that the 100 observations that were excluded because of missing price information are clustered among lower developed countries, thereby removing the variance needed to find the effect of ICT development. The effect is also easily observable in model (3). Without the price variable, "ICT development index" shows a significant and positive relation.

Now the question is, to what extent does our sample reflect the population? Looking back to our sampling strategy, in model (5) of Table 2, we have a model that explained 84% of heterogeneity; so looking for neighbors in the projection of model (5) increases the chance of getting twins that are very comparable for all the factors that we cannot observe and are already somewhat correlated to size measures. This means that the enriched data points contain more valuable information than others from the total population. However, since instead of a totally random sample we are creating twins of providers, our targeted sampling strategy might introduce possible biases. To deal with that bias, we make an assumption that the bias is linear in the modeling domain, i.e., can be captured by linear fixed effects parameters. In Section 7, we further perform additional cross-validations, to check for possible biases our methodology might have introduced.

7 ROBUSTNESS CHECKS

During the course of our analysis, we pointed out a few assumptions that we have made, most notably about the impact of the deviations from the Poisson model due to over-dispersion and about the impact of measurement errors in the abuse data. In this section, we address these two concerns. Regarding the first assumption, we use a simulation study to perform a robustness check on our size estimates. To check the robustness of our method against possible errors in the our phishing blacklists data, we cross validate our results with another, largely independent data source.

7.1 Size Estimates

Assume attack events *are* Poisson and the only structural factor that affects them is the size, i.e., the attack surface. In the absence of perfect information of the true attack surface, it can only be approximated in practice through the size variables we used in Table 2. Now we would like to study: To what extent are deviations from Poisson observable only by using the imperfect size estimations? The precise steps toward estimating new models for phishing abuse counts using imperfect size estimations are as follows:

We generate a *true size* variable that is following normal distribution using mean and standard deviation of variable "Number of hosted domains". We then simulate attack events—*hits*—where the phishing counts follow a Poisson distribution ($Pois(\lambda_{sim})$), where λ_{sim} is the mean number of phishing domains. We then build the simulated size variables used in model (3) of Table 2, namely, "Number of hosted domains," "Number of IP addresses hosting domains," "Number of the assigned IP addresses" by adding random noise ($N(\mu_{f_i}, \sigma_{f_i}), \forall i \in \{1..3\}$) to our true size variable. μ_{f_i} and σ_{f_i} are estimated using the mean and standard deviation of the corresponding explanatory size variables.

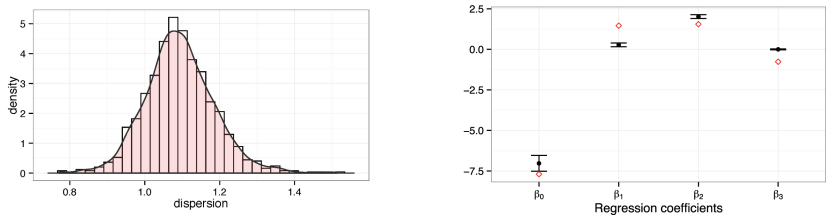


Fig. 6. (Left) Density of dispersion estimates obtained from synthetic Poisson phishing arrival with imperfect size control. (Right) Error bars of the simulated regression coefficients.

We generate 1,000 times the synthetic data representing both the size and the dependent variables, and model them using the GLM regression specified in model (4) (see Table 2). Then we calculate the dispersion parameter for each one of the simulated models using Equation (2). Figure 6(a) shows the density distribution of the dispersion parameter for each of the Poisson models using simulated size measures. The model indicates that the dispersion parameter is on average greater than 1. The dispersion parameter value is far from the dispersion parameter that we obtained using our real dataset in model (4) of Table 2. This, however, is expected, since the real size estimates are far from being perfect and contain several outliers. Moreover, the over-dispersion in simulated size variables indicates that some part of the over-dispersion in model (4) of Table 2—probably not everything—can be attributed to the approximate measurement of the size estimates. Finally, Figure 6(b) displays the coefficients of the 1,000 model fits as error bars. The red diamonds are the coefficients obtained with the full Poisson model. The coefficients follow the same trend as in the model given the over-dispersion, which validates its robustness.

7.2 Phishing Data

Limitations of abuse blacklists, such as comprehensiveness and independence, have been studied at length in prior work [e.g., Metcalf and Spring 2013, 2015]. To study the effect of such limitations on our results, we applied our approach to an alternative dataset: the Cyscon phishing data. These in APWG and Cyscon data have a 13% overlap of unique second-level phishing domains. This provides enough independent observations to corroborate our approach and asses the sensitivity of our results to measurement errors in phishing blacklists.

Similar to before, we model arrival rate of phishing counts using a Poisson GLM with log-link function with size and business model predictors as used in models (4) and (5) of Table 2. The result of the two final models is displayed in Table 4. We then model the statistical twins in the set of rich datapoints $\mathcal{R}$, having two sets of fixed effects for differences between twins and countries. The results are shown in Table 5. Models (1) and (2) contain the same explanatory variables as models (4) and (5) of Table 3. Reassuringly, the resulting estimated coefficients and significance levels for both of the analysis are quite similar to those of the model with APWG data.

8 CONCLUSIONS AND OUTLOOK

The core question of this article is: To what extent are abuse levels determined by the structural properties of providers versus being mainly determined by other factors, including, but not limited to, the security efforts of individual providers? Below, we summarize our findings and discuss the implications of the results.

We reduced errors in the attribution of abuse by empirically studying the population of hosting providers, which are defined based on organizations to whom IP address space is assigned, rather than by routing data and AS ownership. Next, we advanced the existing work that uses simple

Table 4. GLM for Count of Phishing Domains in Cyscon Data for **All Providers**

Response Variable: Count of phishing domains		
Poisson with Log Link Function		
	(1)	(2)
Number of assigned IPs	-0.719^{***}	-0.776^{***}
	(0.011)	(0.012)
Number of IPs hosting domains	1.170^{***}	1.751^{***}
	(0.014)	(0.018)
Number of hosted domains	1.663^{***}	1.115^{***}
	(0.007)	(0.011)
Percentage of domains hosted on shared IPs		0.033^{***}
		(0.001)
Constant	-6.432^{***}	-8.488^{***}
	(0.026)	(0.045)
Observations	45,358	45,358
Log likelihood	−49,763.500	−47,208.470
Akaike Inf. Crit.	99,535.010	94,426.950
Dispersion	20.153	17.444
Pseudo R^2	0.791	0.803

Note: Standard errors in brackets. *p<0.1; **p<0.05; ***p<0.01.

Table 5. GLM for Count of Phishing Domains in Cyscon Data for **Statistical Twins**

Response Variable: Count of phishing domains		
Poisson with Log Link Function		
	(1)	(2)
Price per year	0.0003	-0.012^{***}
	(0.0003)	(0.002)
Popularity index (in thousands)	0.02^{***}	0.1^{***}
	(0.004)	(0.01)
Time in business	-0.048^{***}	0.004
	(0.010)	(0.037)
ICT dev. index		13.610
		($>10^3$)
Wordpress use	7.848^{***}	3.079^{**}
	(0.583)	(1.125)
Pair Fixed-Effect	Yes	Yes
Country Fixed-Effect	No	Yes
Observations	84	82
Log likelihood	−476.818	−145.676
Akaike Inf. Crit.	1,045.635	433.352
Dispersion	20.712	2.993
Fixed-effects pseudo R^2	0.538	0.889
Total pseudo R^2	0.970	0.987

Note: Standard errors in brackets. *p<0.05; **p<0.01; ***p<0.001.

regression analysis and naive normalizations by studying a variety of factors and errors at work that can potentially explain abuse counts. By building several GLM models for phishing abuse counts as the response variable, we demonstrated that a handful of providers' structural properties—such as the number of domain names, number of IP addresses used for web-hosting, and the size of their shared hosting business—can already account for 84% of the variation in phishing counts. These variables are easily measurable at scale and capture the "attack surface" of providers along with aspects of their "business model."

Additionally, we measured the impact of previously unstudied factors, such as price, time-in-business of a provider, and the amount of Wordpress sites per provider. These were collected via a tailored sampling approach and explained a further 77% of the remaining variation in phishing abuse.

Finally, we performed a set of robustness checks on the assumption we made during our analysis. The results of the simulation study performed to check the robustness of our size estimates indicated that coefficients of size variables follow the same trend as in our model (Table 2) given the over-dispersion. To check robustness of our method against observational biases in the phishing data, we cross-validated our results from APWG blacklist data against Cycson phishing blacklist data and observed very similar results.

Our findings suggest that abuse rates for phishing reflect an overall bad harvest, rather than being driven by some rotten apples, i.e., providers that do not care about security. In other words, referring back to the explained and unexplained variance by our models, we observe that structural properties of providers explain the majority of variance in phishing abuse counts, leaving a thin margin for other unmeasured factors including, but not limited to, the security efforts of providers. When structural factors are so dominant in driving abuse, it undermines the common narrative to

call for better security practices of apparently under-performing actors or for even more intrusive interventions, such as sanctions. However, our findings do not limit the action space for policy. Quite the contrary: data-driven policy could try to improve the factors identified as influential, e. g., require higher security standards at providers who host more popular websites.

Our approach enhances more informative comparisons of providers' security performance. In other words, it generates comparative abuse metrics by controlling for the structural differences among providers. Such relative metrics are more suited to evaluate countermeasures than absolute counts or relative counts that generated by naive normalization of one size estimate. Additionally, relative metrics can, in themselves, incentivize better security (Konte et al. 2015; He et al. 2015). In sum: throwing out a few rotten apples might appear more tractable, but producing a better harvest is definitely possible.

Here, we should also acknowledge several limitations of our work. First, our method is geared toward identifying the main explanatory factors in the population of hosting providers. Our conclusions should not be misinterpreted as evidence that there are no misbehaving or negligent hosting providers, only that their impact on the population of phishing incidents is surprisingly limited.

Second, the presence of certain unobserved factors, including security effort and attacker behavior, is a limitation of this work. We have reduced the likelihood of these being major factors in the abuse patterns, as witnessed by the variance explained by the structural properties. We are able to explain 84% of the variance by the structural factors alone—and even more when we take the findings of the statistical twins sample into account. The remaining unexplained variance, which is the combination of provider's security effort, attacker behavior, and measurement errors, suggests that the impact of provider's security effort should be limited. That being said, the only way to determine the precise impact of security effort of providers on abuse levels is by directly measuring it. This is a challenge for future work, which might deploy provider surveys or experiments with abuse notifications and take-down speed (Li et al. 2016; Cetin et al. 2015).

Third, certain structural factors might indirectly capture some information about security efforts. One could argue, for example, that the pricing model chosen by a provider might also contain a signal about the amount of resources available for security. However, the fact that 84% of the variance is explained by purely technical structural properties, unrelated to price, suggests that also this impact is limited. Only a more direct observation of security effort can establish how it is related to price.

A final limitation is that our empirical evidence is specific to phishing. Our modeling approach is agnostic to the type of abuse, however. The independent variables and model design are not specific to phishing. Future work can use our approach to identify the impact of these structural and business model factors on other types of abuse in the hosting market. For some sources, like drive-by-download sites, we expect similar patterns. For other, more idiosyncratic types of abuse, like long-living botnet C&C servers, we might expect different patterns. There, we might indeed find that rotten apples drive the abuse rates, rather than a bad overall harvest.

ACKNOWLEDGMENTS

The authors thank Paul Vixie and Eric Ziegast from Farsight Security for sharing DNSDB and Thorsten Kraft from Cyscon for providing data on phishing websites.

REFERENCES

Greg Aaron and Rod Rasmussen. 2015a. Anti-phishing working group (APWG) global phishing survey: Trends and domain name use in 2H2014. Retrieved from http://internetidentity.com/wp-content/uploads/2015/05/APWG_Global_Phishing_Report_2H_2014.pdf.

Greg Aaron and Rod Rasmussen. 2015b. Global phishing survey: Trends and domain name use in 1H2014. Retrieved from http://docs.apwg.org/reports/APWG_Global_Phishing_Report_1H_2014.pdf.

Anti-Phishing Working Group. 2016. Retrieved from http://www.antiphishing.org.

Hadi Asghari, Michael Ciere, and Michel J. G. Van Eeten. 2015a. Post-mortem of a zombie: Conficker cleanup after six years. In *Proceedings of the 24th USENIX Security Symposium (USENIXSecurity'15)*. 1–16.

Hadi Asghari, Michel J. G. van Eeten, and Johannes M. Bauer. 2015b. Economics of fighting botnets: Lessons from a decade of mitigation. *IEEE Secur. Priv.* 5 (2015), 16–23.

Leyla Bilge, Engin Kirda, Christopher Kruegel, and Marco Balduzzi. 2011. EXPOSURE: Finding malicious domains using passive DNS analysis. In *Proceedings of the Network & Distributed System Security Symposium (NDSS'11)*. The Internet Society, 1–17.

Xue Cai, John Heidemann, Balachander Krishnamurthy, and Walter Willinger. 2010. Towards an AS-to-organization map. In *Proceedings of the 10th Internet Measurement Conference (IMC'10)*. ACM, 199–205.

A. Colin Cameron and Pravin K. Trivedi. 1990. Regression-based tests for overdispersion in the poisson model. *J. Econ.* 46, 3 (1990), 347–364.

A. Colin Cameron and Pravin K. Trivedi. 2013. *Regression Analysis of Count Data*, vol. 53. Cambridge University Press.

Davide Canali, Davide Balzarotti, and Aurélien Francillon. 2013. The role of web hosting providers in detecting compromised websites. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. 177–188.

Orcun Cetin, Mohammad Hanif Jhaveri, Carlos Gañán, Michel van Eeten, and Tyler Moore. 2015. Understanding the role of sender reputation in abuse reporting and cleanup. In *Proceedings of the 14th Workshop on the Economics of Information Security (WEIS'15)*. 1–15.

Richard Clayton, Tyler Moore, and Nicolas Christin. 2015. Concentrating correctly on cybercrime concentration. In *Proceedings of the 14th Annual Workshop on the Economics of Information Security (WEIS'15)*. 1–16.

M. Patrick Collins, Timothy J. Shimeall, Sidney Faber, Jeff Janies, Rhiannon Weaver, Markus De Shon, and Joseph Kadane. 2007. Using uncleanliness to predict future botnet addresses. In *Proceedings of the 7th Internet Measurement Conference (IMC'07)*. ACM, 93–104.

Jakub Czyz, Michael Kallitsis, Manaf Gharaibeh, Christos Papadopoulos, Michael Bailey, and Manish Karir. 2014. Taming the 800 pound gorilla: The rise and decline of NTP DDoS attacks. In *Proceedings of the 14th Internet Measurement Conference (IMC'14)*. ACM, 435–448.

X. Dimitropoulos, D. Krioukov, G. Riley, and K. Claffy. 2006. Revealing the autonomous system taxonomy: The machine learning approach. In *Proceedings of the Passive and Active Network Measurement Workshop (PAM'06)*. 91–100.

DNS Database (DNSDB). 2016. Farsight Security. Retrieved from https://www.dnsdb.info.

Dutch Hosting Provider Association. 2013. Nederland paradijs voor internet criminelen? Retrieved from https://www.dhpa.nl/nederland-paradijs-voor-internet-criminelen.html.

Kathryn Elliott. 2008. Who, what, where, when, and why of WHOIS: Privacy and accuracy concerns of the WHOIS database. *SMU Sci. Technol. Law Rev.* 12 (2008), 141.

Farsight Security. 2016. Security Information Exchange. Retrieved from https://www.farsightsecurity.com.

Vaibhav Garg and L. Jean Camp. 2013. Macroeconomic analysis of malware. In *Proceedings of the Network & Distributed System Security Symposium (NDSS'13)*. The Internet Society, 1–3.

Cyscon GmbH. 2016. Cyscon Security - PhishKiller. Retrieved from http://www.cyscon.de.

Max Goncharov. 2015. Criminal Hideouts for Lease: Bulletproof Hosting Services. Retrieved from http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-criminal-hideouts-for-lease.pdf.

Shuang Hao, Nick Feamster, and Ramakant Pandrangi. 2011. Monitoring the initial DNS behavior of malicious domains. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*. ACM, 269–278.

Shuang Hao, Matthew Thomas, Vern Paxson, Nick Feamster, Christian Kreibich, Chris Grier, and Scott Hollenbeck. 2013. Understanding the domain registration behavior of spammers. In *Proceedings of 13th Internet Measurement Conference (IMC'13)*. ACM, 63–76.

Shu He, Gene Moo Lee, John S. Quarterman, Quarterman Creations, and Andrew B. Whinston. 2015. Cybersecurity policies design and evaluation: Evidence from a large-scale randomized field experiment. In *Proceedings of the 14th Annual Workshop on the Economics of Information Security (WEIS'15)*. 1–50.

Harald Heinzl and Martina Mittlböck. 2003. Pseudo R-squared measures for poisson regression models with over-or underdispersion. *Comput. Stat. Data Anal.* 44, 1 (2003), 253–271.

HostExploit. 2017. World Hosts Report. Retrieved from http://hostexploit.com.

International Telecommunication Union (ITU). 2014. Measuring the Information Society Report 2014. Retrieved from https://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2014/MIS2014_without_Annex_4.pdf.

Aaron Kleiner, Paul Nicholas, and Kevin Sullivan. 2013. Linking cybersecurity policy and performance. *Microsoft Trust. Comput.* 1, 1 (2013), 1–20.

Maria Konte, Roberto Perdisci, and Nick Feamster. 2015. ASwatch: An AS reputation system to expose bulletproof hosting ASes. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication (SIGCOMM'15)*. ACM, 625–638. DOI: http://dx.doi.org/10.1145/2785956.2787494

Marc Kührer, Christian Rossow, and Thorsten Holz. 2014. Paint it black: Evaluating the effectiveness of malware blacklists. In *Research in Attacks, Intrusions and Defenses*. Springer, 1–21.

Kirill Levchenko, Andreas Pitsillidis, Neha Chachra, Brandon Enright, Márk Félegyházi, Chris Grier, Tristan Halvorson, Chris Kanich, Christian Kreibich, He Liu et al. 2011. Click trajectories: End-to-end analysis of the spam value chain. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'11)*. IEEE, 431–446.

Frank Li, Grant Ho, Eric Kuan, Yuan Niu, Lucas Ballard, Kurt Thomas, Elie Bursztein, and Vern Paxson. 2016. Remedying web hijacking: Notification effectiveness and webmaster comprehension. In *Proceedings of the 25th International Conference on the World Wide Web (WWW'16)*. 1009–1019.

He Liu, Kirill Levchenko, Márk Félegyházi, Christian Kreibich, Gregor Maier, Geoffrey M. Voelker, and Stefan Savage. 2011. On the effects of registrar-level intervention. In *Proceedings of the Conference on Large-scale Exploits and Emergent Threats (LEET'11)*.

Suqi Liu, Ian Foster, Stefan Savage, Geoffrey M. Voelker, and Lawrence K. Saul. 2015. Who is .com? Learning to parse WHOIS records. In *Proceedings of the 15th Internet Measurement Conference (IMC'15)*. ACM, 369–380.

Yang Liu, Armin Sarabi, Jing Zhang, Parinaz Naghizadeh, Manish Karir, Michael Bailey, and Mingyan Liu. 2015. Cloudy with a chance of breach: Forecasting cyber security incidents. In *Proceedings of the 24th USENIX Security Symposium (USENIXSecurity'15)*. 1009–1024.

M3AAWG. 2015. Anti-Abuse Best Common Practices for Hosting and Cloud Service Providers. Retrieved from https://www.m3aawg.org/sites/maawg/files/news/M3AAWG_Hosting_Abuse_BCPs-2015-03.pdf.

MaxMind. 2016. IP Geolocation Databases. Retrieved from https://www.maxmind.com.

McAfee Intel Security. 2013. Botnet Control Servers Span the Globe. Retrieved from https://blogs.mcafee.com/mcafee-labs/botnet-control-servers-span-the-globe.

Leigh Metcalf and Jonathan M. Spring. 2013. *Everything You Wanted to Know About Blacklists But Were Afraid to Ask*. Technical Report. CERT Network Situational Awareness Group.

Martina Mittlböck. 2002. Calculating adjusted R(2) measures for poisson regression models. *Comput. Methods Programs Biomed.* 68, 3 (2002), 205–214.

Nederlandse Omroep Stichting. 2013. Nederland paradijs cybercriminelen. Retrieved from http://nos.nl/artikel/469969-nederland-paradijs-cybercriminelen.html.

Nick Nikiforakis, Wouter Joosen, and Martin Johns. 2011. Abusing locality in shared web hosting. In *Proceedings of the Fourth European Workshop on System Security*. ACM, 2.

Arman Noroozian, Maciej Korczyński, Samaneh Tajalizadehkhoob, and Michel van Eeten. 2015. Developing security reputation metrics for hosting providers. In *Proceedings of the 8th USENIX Workshop on Cyber Security Experimentation and Test (CSET'15)*. 1–8.

Andreas Pitsillidis, Chris Kanich, Geoffrey M. Voelker, Kirill Levchenko, and Stefan Savage. 2012. Taster's choice: A comparative analysis of spam feeds. In *Proceedings of the 12th Internet Measurement Conference (IMC'12)*. ACM, 427–440.

Anirudh Ramachandran and Nick Feamster. 2006. Understanding the network-level behavior of spammers. *ACM SIGCOMM Comput. Commun. Rev.* 36, 4 (2006), 291–302.

Angelo P. E. Rosiello, Engin Kirda, Christopher Kruegel, and Fabrizio Ferrandi. 2007. A layout-similarity-based approach for detecting phishing pages. In *Proceedings of the 3rd International SecureComm Conference*. IEEE, 454–463.

Craig A. Shue, Andrew J. Kalafut, and Minaxi Gupta. 2012. Abnormally malicious autonomous systems and their internet connectivity. *IEEE/ACM Trans. Netw.* 20, 1 (2012), 220–230.

Kyle Soska and Nicolas Christin. 2014. Automatically detecting vulnerable websites before they turn malicious. In *Proceedings of the 23rd USENIX Security Symposium (USENIXSecurity'14)*. USENIX, 625–640.

Brett Stone-Gross, Christopher Kruegel, Kevin Almeroth, Andreas Moser, and Engin Kirda. 2009. Fire: Finding rogue networks. In *Proceedings of the Computer Security Applications Conference*. IEEE, 231–240.

S. Tajalizadehkhoob, M. Korczynski, A. Noroozian, C. Ganán, and M. van Eeten. 2016. Apples, oranges and hosting providers: Heterogeneity and security in the hosting market. In *Proceedings of the Network Operations and Management Symposium (NOMS'16)*. IEEE/IFIP, 289–297. DOI: http://dx.doi.org/10.1109/NOMS.2016.7502824

M. Vasek, J. Wadleigh, and T. Moore. 2016. Hacking is not random: A case-control study of webserver-compromise risk. *IEEE Trans. Depend. Secure Comput.* 13, 2 (2016), 206–219.

Christoph Wagner, Jérôme François, Radu State, Alexandre Dulaunoy, Thomas Engel, and Gilles Massen. 2013. ASMATRA: Ranking ASs providing transit service to malware hosters. In *Proceedings of the Conference on Integrated Management (IM'13)*. IFIP/IEEE, 260–268.

Web-Archive. 2016. Internet Archive. Retrieved from http://archive.org/web.

Colin Whittaker, Brian Ryner, and Marria Nazif. 2010. Large-scale automatic classification of phishing pages. In *Proceedings of the Network & Distributed System Security Symposium (NDSS'10)*. The Internet Society, 1–5.

WPScan Team. 2016. WordPress Vulnerability Scanner. Retrieved from http://wpscan.org.

Jing Zhang, Zakir Durumeric, Michael Bailey, Mingyan Liu, and Manish Karir. 2014. On the mismanagement and maliciousness of networks. In *Proceedings of the Network & Distributed System Security Symposium (NDSS'14)*. The Internet Society, 1–12.

Received November 2016; revised April 2017; accepted June 2017