Inferring the Security Performance of Providers from Noisy and Heterogenous Abuse Datasets

Arman Noroozian ✉, Michael Ciere, Maciej Korczyński, Samaneh Tajalizadehkhoob, and Michel van Eeten

Delft University of Technology

Abstract

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. 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. We present the first comprehensive approach to measure defender security performance from a combination of heterogeneous abuse datasets, taking all of these issues into account. We present a causal model of incidents, test for biases across seven abuse datasets and then propose a new modeling approach. Using Item Response Theory, we estimate the security performance of providers as a latent, unobservable trait. The approach also allows us to quantify the uncertainty of the performance estimates. Despite the uncertainties, we demonstrate the effectiveness of the approach by using the security performance estimates to predict a large portion of the variance in the abuse counts observed in independent datasets, after controlling for various exposure effects such as the size and business type of the providers.

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 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 [1], Internet Service Providers (ISPs) [2, 3], countries [4, 5], organizations [6], payment providers [7], registrars [8], registries [9], 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 [10, 11].

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 [12, 13], to identify bad providers [14, 15], and to study the effectiveness of countermeasures [16].

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.

Second, abuse datasets are highly heterogeneous. They are very different in size. Some sets observe one or two order of magnitude more events than others. They also have only very little overlap among them [17]. 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., provider). Some providers might be more susceptible to certain types of abuse, but less to others.

A third problem is the lack of completeness. 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 [11, 14]. 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. 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. We will discuss this in more detail in the section on related work. We present the first comprehensive approach to measure defender security performance from a combination of heterogeneous abuse datasets. We apply the approach to the hosting sector, which is associated with a large portion of all observed abuse events. We first present a causal model to explain abuse rates in provider networks. We then map the providers in the hosting market. Second, we study potential biases in the distribution of abuse data over providers. Next, we 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, we test the reliability of the performance metric.

Our contributions are as follows:

- We formalize a causal model in order to systematically disentangle the different factors at work in abuse data. It provides a basis for modeling security economics questions based on incident data.
- We show that a combination of 7 abuse datasets covers 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, we demonstrate that all providers, small and large, have equal probability of showing up in abuse data, once we control for their exposure.
- We present a novel statistical approach – based on Item Response Theory – 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, we 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.

The overall goal of our study is to enable better measurement of security performance from 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 we will first discuss the causal abuse framework which forms the background of our work in Section 2. We then provide an overview of our data in Section 3. To explore the bias in our abuse data, we map the hosting provider market using several other data sources in Section 4 and find no evidence of observation bias using simple simulations of attacks across the hosting market in Section 5. We then move on the construct our IRT model and motivate our approach in Section 6, then provide the specification of the model in Section 7 and estimate the security performance of the hosting providers in Section 8. The robustness and predictive power of our security performance estimates are explored in Section 9. Finally we provide an overview of the related work, and studies on which our work builds in Section 10 and finally discuss the implication of our work and conclude in Section 11.

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 [18], provider properties and business models [11], or the effect of interventions [4, 16].

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 [11]. In this paper, we focus on the causal factors. Figure 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 the attack. 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.

The yellow ovals in Figure 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 [11, 19]. 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 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 [13, 14, 18]. Price is another example. Cheap or free services are more prone to be abused by miscreants, leaving their providers more exposed [11].

With this causal model in hand, we can more precisely articulate the core idea of

[1] The authors 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.

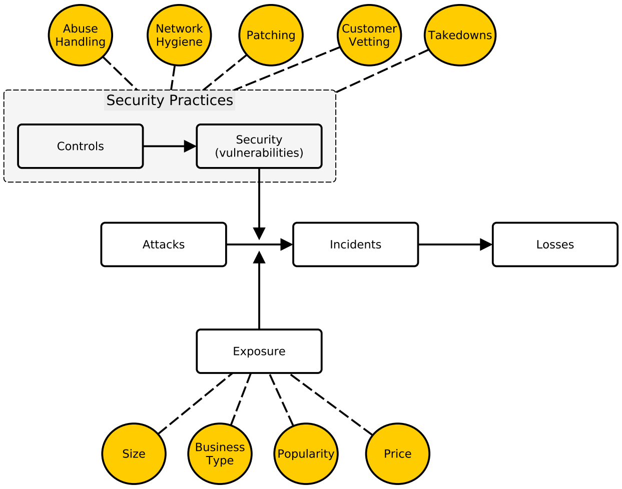


Figure 1: Causal Model Incidents

this paper. We 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. We would like to test to what extent performance can be estimated reliably as a latent factor that is driving the abuse count.

The model illustrates that our approach assumes we can control for exposure and attacks. The former we will include in our models via a number of indicators, which we will collect for the whole population of hosting providers. The latter we cannot observe directly and we will include it as a random variable. In other words: we assume that attacks are randomly distributed across the attack surface. In section 4, we will test this assumption via a simple simulation. To the extent that this assumption does not hold, it will increase the size of the error term of our model, i.e., leave more variance in the abuse data unexplained.

3 Data

3.1 Abuse Data

Since we are interested in hosting providers, we use seven data feeds that include incidents typical for hosting services: phishing and malware-related abuse. The malware data provided to us by Stopbadware Data Sharing Program, contains feeds from a number of volunteer companies and research institutions for the entire duration of 2015 [20]. The dataset contains URLs and IP addresses associated with 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:

Table 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	1,580 - 7,208 ** (ranging between)	11,976 - 376,561 ** (ranging between)	Malware	Stopbadware DSP
SBW2	2015			Malware	Stopbadware DSP
SBW3	2015			Malware	Stopbadware DSP
SBW4	2015			Malware	Stopbadware DSP
SBW5	2015			Malware	Stopbadware DSP

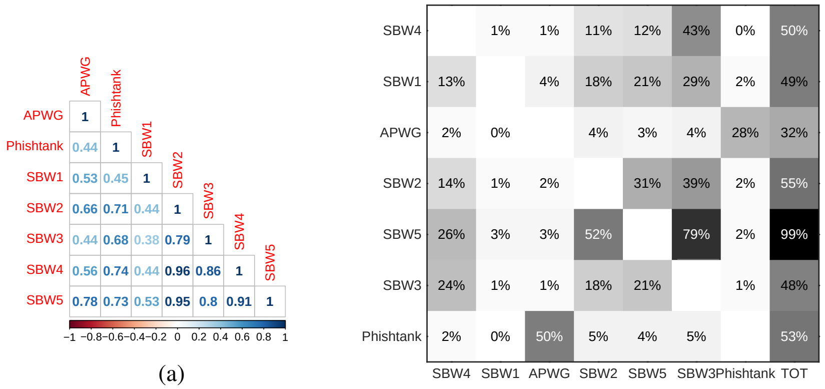
** Due to the terms of the data sharing agreement, we only report aggregated ranges for SBW data

Anti-Phishing Working Group (APWG) [21] and Phishtank [22]. Both datasets contain IP addresses, fully qualified domain names and URLs associated with phishing. Table 1 provides a summary of our abuse feeds, the abused organizations and the number of incidents they had according to each feed in 2015.

For each dataset, we 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. The metric we define as event per provider is the number of unique (2nd-level-domain, IP-Address) pairs recorded per provider in every abuse feed.

Most concentration metrics choose Autonomous Systems (ASes) as the unit of analysis [13, 14, 23] and associate events with AS owners based on the BGP prefix announcements for each AS. The AS owner, however, often merely routes the traffic for the IP address and has no administrative responsibility for it. In earlier work [24], we developed an 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, we select the hosting providers through a series of steps which we explain below in Section 4.

Figure 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 of abuse, may be weakly correlated with each other. The correlation 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 2b on the other hand illustrates the overlap between our abuse feeds in terms of what percentage of 2nd-level-domains 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.



(a)

(b)

Figure 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.

3.2 Hosting Data

To construct a mapping of the hosting provider market, we use several data sources and build on techniques used in previous work [14, 24]. Our mapping approach to identify hosting provider organizations is based on (i) IP ownership data from Maxmind's WHOIS API [25] and (ii) passive DNS data from DNSDB [26] generously provided to us by Farsight Security. The passive DNS data contains fully-qualified domain names 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 organizations that we can use as proxies for their exposure (see Figure 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 2nd-level domains (2LDs) hosted by the organization, (iv) the number of IP addresses that are associated with at least 10 2LDs (proxy for shared hosting), and (v) the number of 2LDs on shared IPs hosted by an organization.

4 Hosting Provider Market

Our 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. We then use passive DNS data to construct the remaining structural properties (see Section 3.2) based on what has been passively observed in DNS traffic over the duration of 2015.

We define hosting providers as the subset of organizations for which we have observed at least 30 2LDs, a low threshold to minimize false negatives. All others are considered non-hosting organizations. Figure 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.

A comparison of the distributions of 'all' organizations and those 'observed in DNSDB' (purple vs green) demonstrates that DNSDB provides a reasonably unbiased view of organizations, and thus providers, as the shapes of the two distributions closely follow the same pattern, especially for organizations that own 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 [9].

We 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 of the hosting provider market. Note from the distribution of hosting providers in Figure 3 (blue) that the bulk of these providers have been allocated between 300 to 10,000 IP addresses.

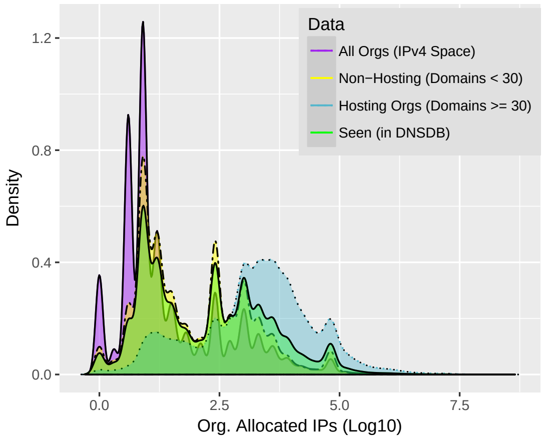
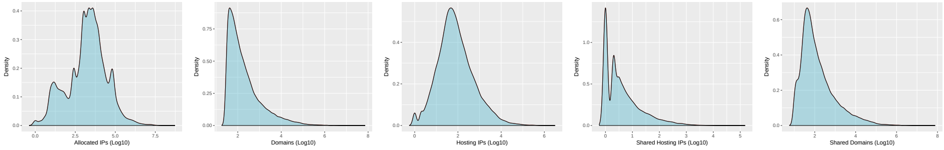


Figure 3: Distributions of Allocated IP Space

Given our definition of the hosting providers and the passive DNS data, we can empirically construct a picture of the aforementioned 'exposure properties' of each hosting provider. Figure 4 plots the distributions of these properties for all hosting providers.

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 (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 services for each provider. Figure 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 residing on shared hosting. In other words, the provider is increasingly dependent on shared hosting as its main business model in webhosting.

For brevity, we will not go into more detail about the provider mapping and instead



(a) Allocated IPs (b) Domains (c) Hosting IPs (d) Shared IPs (e) Shared Doms

Figure 4: Distributions of hosting providers over exposure properties

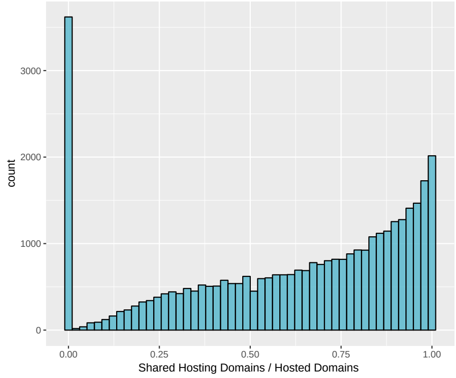


Figure 5: Shared vs dedicated hosting

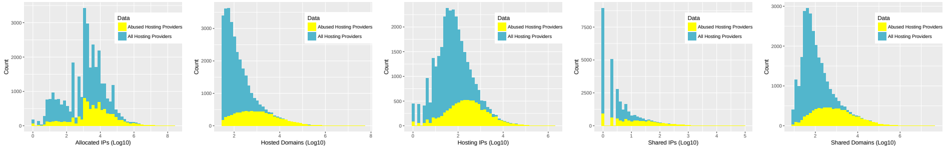
refer the reader to [24] for a more in depth analysis of the market.

5 Exploring Observation Bias in Abuse Data

We now explore how our 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 contains incidents. 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 6 shows histograms for each of the exposure properties. Each histogram 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 these attacks that the probability of incurring a single abuse 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. We test this explanation via a simple simulation.



(a) Allocated IPs (b) Domains (c) Hosting IPs (d) Shared IPs (e) Shared Doms

Figure 6: Distributions of hosting providers with (yellow) and without (blue) observed abuse events, over different exposure properties

Understanding the potential observation biases in the abuse data is a key consideration in constructing abuse concentration metrics as previous research points out [10, 14]. One way to identify bias is to compare the datasets against other sources of abuse data. Kührer et al. [27] compared abuse blacklists against each other and against data they collected themselves. In a way, we have done something similar by using seven datasets. They all display the same pattern.

While such comparisons are helpful, the other datasets are not ground truth. They are 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. We therefore complement our 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?

We assume that attackers attack domains at random with a fixed probability. Note that our datasets (see Table 1) mainly capture cybercrime that involves domain names. Therefore, the number of domains of a provider is a useful proxy for the 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 $1-(1-p)^n$. We obtain n from the exposure properties of the provider in our hosting provider mapping. Using a maximum likelihood estimator, we estimate p from our observed abuse data which results in a value of $p = 0.0025$. Given this estimated probability, Figure 7 illustrates a 'separation plot' [28] of the predicted and observed abuse status of all hosting providers.

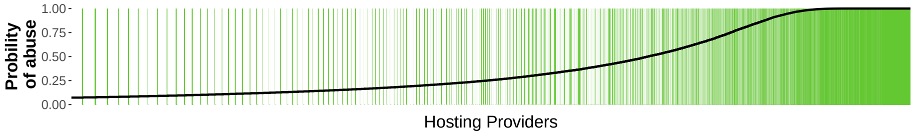
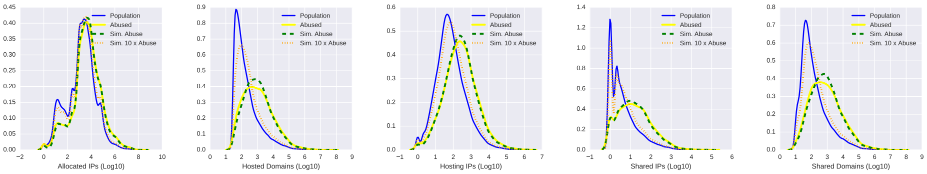


Figure 7: Separation plot of predicted versus observed abuse of hosting providers

The plot demonstrates the degree to which the calculated probability of abuse per domain agrees with the actual observed abuse data. Here, the horizontal axis, and the trend line respectively illustrate all hosting providers and the probability with which we predict they will be abused, sorted in an increasing order. A green tinted thin line represents a provider for which an abuse event has been observed in our abuse feeds. Darker green areas indicate high density of providers with abuse, light green or white areas indicate few or no such providers. The concentration of abused providers towards the right side of the plot illustrates the large degree to which our estimation results and the observed abuse data in Figure 6 are consistent. Figure 7 demonstrates that our assumption regarding the abuse generation process is reasonable.

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



(a) Allocated IPs (b) Domains (c) Hosting IPs (d) Shared IPs (e) Shared Doms

Figure 8: Non-biased distribution of abuse over population of hosting providers

We compare the distributions of providers over the different exposure properties in Figure 8. The first simulation, generating the same number of events as in our empirical data, produces a distribution that is highly similar with that empirical 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 finding is this: 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.

These results suggest that patterns observed in Figure 6 are not the result of observation bias, but rather of attacker dynamics and the random nature of the attack generation process. The simulation also provides support for a modeling decision that we 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.

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 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. We can then try to infer it 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 the empirical distribution of abuse events over the hosting market by modeling it as random process over the attack surface, as measured by the exposure indicators. This also suggests that our exposure indicators capture an important portion of the exposure factor. A more precise test was conducted by Tajalizadehkhoob et al.[11]. 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. We will test whether our assumptions indeed hold by testing the predictive power of the estimated security performance: what portion of the remaining variance can it explain by providers' security performance, after having controlled for exposure effects. Before we get to that step, though, we first discuss the statistical approach we propose: estimating a latent variable for each provider through a model based on Item Response Theory. 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. Earlier work 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., [11, 29]. 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., [14] uses 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., [1, 18]). 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 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).

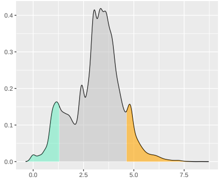
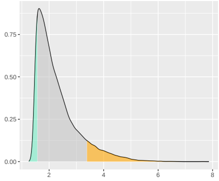
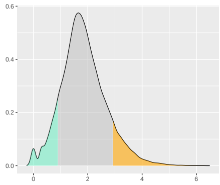
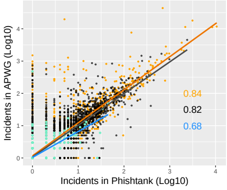
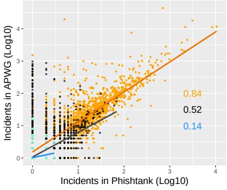
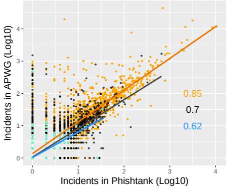
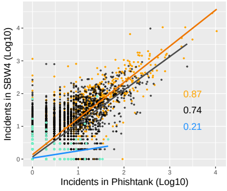
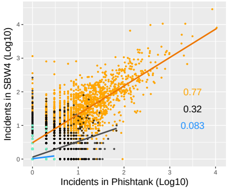
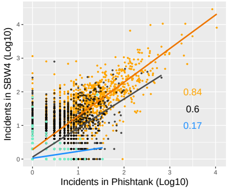
Properties	**Owned IPs**	**Hosted Domains**	**Hosting IPs**
Empirical Distribution			
	⇕	⇕	⇕
APWG / Phish			
SBW4/ *Phish*			
Note:		Reported numbers in plots represent correlation strength	

Table 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. We compare 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 population we inspect as indicated by how strongly the different data points correlate.

To meaningfully capture the different signals within the abuse data and to overcome the aforementioned issues, we apply techniques from Item Response theory [30, 31] 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.

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 [32]. 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 she 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 our abuse feeds as an indicator of how good or bad each student performed in an exam which consists of several questions, which correspond 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.

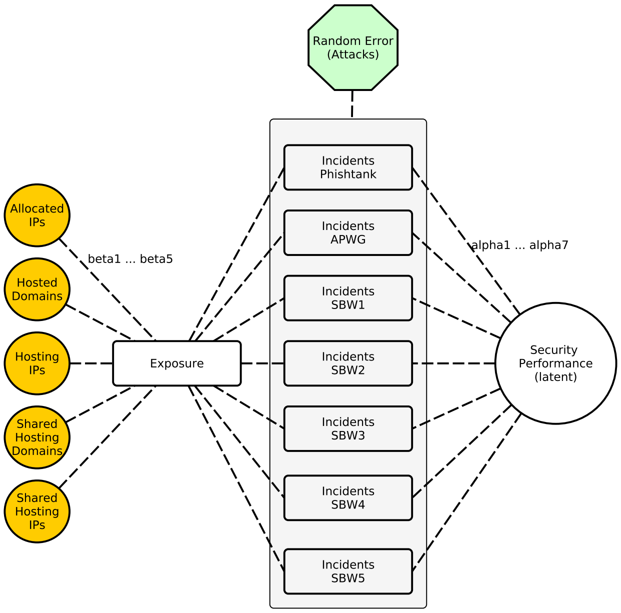


Figure 9: IRT model - Jointly explaining variation in incident count for all abuse feeds

The model is graphically illustrated in Figure 9. For every abuse feed $j = 1, \dots, k$

and for every provider $n = 1, ..., N$, the abuse count Y_{nj} follows a Poisson distribution

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

with

$$\ln(\lambda_{nj}) = \ln(\mathrm{E}(Y_{nj}|\theta_n, \mathbf{x}_n)) = \gamma_j + \mathbf{x}_n^T \boldsymbol{\beta} - \alpha_j \theta_n. \tag{1}$$

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 mitigating 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.

8 Estimation Results

To infer the security performance of providers from abuse data, we input the incident numbers from all abuse feeds into our model and estimate the parameters of our model (see Equation 1) using MCMC simulation. The model uses the exposure related variables (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.

We performed full Bayesian inference of the model parameters and the latent variables by means of Markov Chain Monte Carlo (MCMC) sampling [33]. We used weakly-informative prior distributions

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

reflecting our relative ignorance of their true values. MCMC sampling was carried out using Stan [34] with the `rstan` R interface. We ran 4 chains for 1500 iterations each, with 750 warmup samples. This resulted in a total of 3000 MCMC samples.

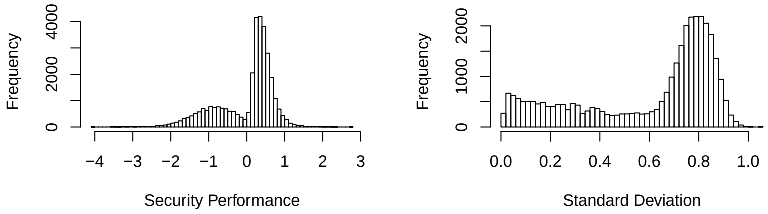
Table 3: IRT model parameter values for all abuse feeds

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

The MCMC algorithm converges towards the parameter values summarized in Table 3 with $\hat{R}$ values close to 1 which indicate convergence of the sampling algorithm [35]. 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 in 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.

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 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 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 values is low. We explain why this large standard deviation occurs shortly here after. Figure 11 illustrates the security performances of all organizations as the posterior mean of the latent variable represented by black dots along side the 95% credible interval of the latent variable values. An orange color 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.



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

Figure 10: Distributions of security performance (latent variable)

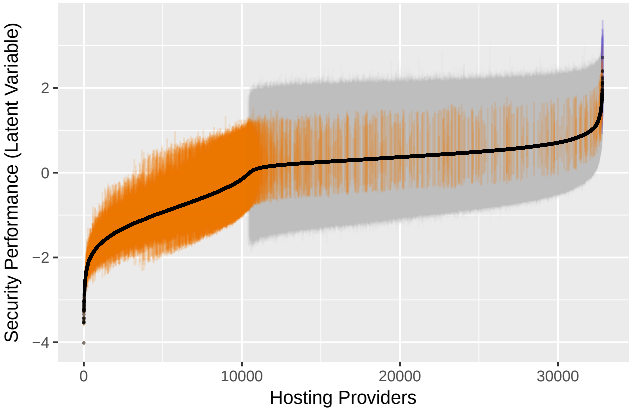


Figure 11: Security performance with 95% credible interval band

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 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 their lack of abuse is due to their small exposure or due to their 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 2 illustrates, for a certain selection of providers, abuse feeds show very different incident counts. These cases, however, all have a standard deviation smaller than 0.5.

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

9 Robustness and Predictive Power

Given our measured 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, we 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 assumed to follow a Poisson distribution of the same form as described in Equation 1. To test the predictive power of our approach, we measure the security performance value repeatedly, each time leaving out one of the abuse feeds. We then use the measured security performance to explain the variance in incident counts in the independent abuse feed. This way, we cross-validate our results and can examine the predictive power of the calculated security performance values. Table 4 shows how different models for the SBW1 dataset explain the variation, where security performance was measured from the other abuse datasets.

Model (1) is a baseline model which only includes a constant 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.

The coefficients for the explanatory variables in the model can be interpreted as follows. We use model (4) as an example, 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).

Table 4: Poisson GLM regression with Log 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 (latent variable)				-1.84^{***} (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	

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 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 calculations. This suggests that the variable is able to capture the latent factor of security performance 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 in the dependent variable.

Table 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 5, is remarkably high for such noisy data on such a multicausal phenomenon.

Two feeds stand out. 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.

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 [18]. 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 it correlates to abuse data [36]. 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 [37]. Vasek et al. combine abuse data sources to identify risk factors for webserver compromise [19]. 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 [1], Internet Service Providers (ISPs) [2, 3], Autonomous Systems [14, 23, 38], countries [4, 5, 29], organizations [6], payment providers [7], registrars [8], registries [9], 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 [10, 11]. Our work 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. Our work is most closely related to [11] in which Tajalizadehkhoob et al. propose analytical models to explain abuse concentrations based on exposure. We build upon this work with a different modeling approach, based on IRT. We also build upon [24] to construct a mapping of the hosting provider market and explore the issue of bias in our 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 [39]. 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 our research is the wide range of problems associated with incident and abuse data, on which a lot of security research is based. Noroozian et al. systematically walk through some of the difficulties associated with creating operator benchmarks based on multiple data sources [14]. Clayton et al. highlight considerations that need to be made before intervening based on abuse concentration metrics [10] 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 [27] 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 [40]. Metcalf and Spring compare the contents of 25 different blacklists and surprisingly find very little overlap between the contents of the blacklists [17]. Our work takes such issues into account and carefully explores the bias in our various data sources to ensure minimal effects on our results.

11 Discussion and Conclusions

The success of many industry and government initiatives to combat cybercrime relies on the ability to empirically track the efforts and progress of various market players. Abuse data is a critical resource in that endeavor, but also a rather unruly one. This study 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 we present a more comprehensive approach that takes all of them into account. We apply the approach to the hosting sector, which is associated with a large portion of all observed abuse events.

We have presented a causal model for the generation of abuse data that is implicitly behind much of the empirical research. We 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 our approach is a modeling approach based on Item Response Theory, which estimates security performance of hosting providers as a unobservable latent variable from an array of abuse datasets. The Bayesian nature of our approach also means that we can quantify the 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. We test the robustness of our approach via out-of-sample predictions. We find that our 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 our approach. Due to the noisy nature of the abuse data and the limitations of our 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. We intend to undertake this in future work.

In sum, we 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 practices and government oversight.

Acknowledgements

This work has been funded by NWO under Pr. Nr. CYBSEC.12.003 / 628.001.003, and SIDN (www.sidn.nl). We would like to thank Dr. Paul Vixie from Farsight Security for providing the DNSDB data which has been used in this study, StopBadware, the Stopbadware DSP program and their data contributors (ESET, Fortinet, Internet Identity, Sophos) including their anonymous feed contributors, in addition to APWG and Phishtank for providing access to their abuse data.

References

[1] Brett Stone-Gross, Christopher Kruegel, Kevin Almeroth, Andreas Moser, and Engin Kirda. "FIRE: FInding Rogue nEtworks". In: *ACSAC*. 2009, pp. 231–240.

[2] M Van Eeten, JM Bauer, and H Asghari. "The role of internet service providers in botnet mitigation an empirical analysis based on spam data". In: *WEIS*. 2010.

[3] Giovane C M Moura, Ramin Sadre, and Aiko Pras. "Bad neighborhoods on the internet". In: *IEEE Communications Magazine* 52.7 (2014), pp. 132–139.

[4] Hadi Asghari, Michael Ciere, and Michel J G Van Eeten. "Post-Mortem of a Zombie: Conficker Cleanup After Six Years". In: *USENIX Security* (2015).

[5] Fanny Lalonde Levesque, Jose M. Fernandez, Anil Somayaji, and Dennis Batchelder. "National-level risk assessment : A multi-country study of malware infections". In: *Proc. of WEIS* (2016), pp. 1–30.

[6] 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: *24th USENIX Security Symposium (USENIX Security 15)* (2015), pp. 1009–1024.

[7] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Felegyhazi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, D. McCoy, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage. "Click Trajectories: End-to-End Analysis of the Spam Value Chain". English. In: *2011 IEEE Symposium on Security and Privacy*. IEEE, 2011, pp. 431–446.

[8] 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 Large-scale exploits and emergent threats*. 2011.

[9] 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: *Proc. 2nd IEEE European Symposium on Security and Privacy (EuroS&P’17)*. 2017.

[10] Richard Clayton, Tyler Moore, and Nicolas Christin. “Concentrating Correctly on Cybercrime Concentration”. In: *WEIS*. 2015, pp. 1–16.

[11] 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*. 2017. URL: http://arxiv.org/abs/1702.01624.

[12] 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.

[13] *HostExploit*. URL: http://hostexploit.com/.

[14] Arman Noroozian, Maciej Korczynski, Samaneh Tajalizadehkhoob, and Michel van Eeten. “Developing Security Reputation Metrics for Hosting Providers”. In: *USENIX Workshop on Cyber Security Experimentation and Test (USENIX CSET’15)* (2015).

[15] Andrew J Kalafut, Craig A Shue, and Minaxi Gupta. “Malicious Hubs: Detecting Abnormally Malicious Autonomous Systems”. In: *2010 Proceedings IEEE INFOCOM*. IEEE, 2010, pp. 1–5.

[16] Benjamin Edwards, Steven Hofmeyr, Stephanie Forrest, and Michel van Eeten. “Analyzing and Modeling Longitudinal Security Data: Promise and Pitfalls”. In: *Proceedings of the 31st Annual Computer Security Applications Conference on - ACSAC 2015*. ACM Press, 2015, pp. 391–400.

[17] 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.

[18] J Zhang, Z Durumeric, and M Bailey. “On the Mismanagement and Maliciousness of Networks”. In: *NDSS*. 2014.

[19] 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.

[20] *StopBadware*. URL: https://www.stopbadware.org/data-sharing.

[21] *Anti Phishing Working Group - APWG*. URL: http://www.antiphishing.org.

[22] *Phishtank*. URL: https://www.phishtank.com/index.php.

[23] Maria Konte, Roberto Perdisci, and Nick Feamster. “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 Press, 2015, pp. 625–638.

[24] 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 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2016, pp. 289–297.

[25] *Maxmind GeoIP2 DB*. URL: https://www.maxmind.com/en/geoip2-isp-database.

[26] Farsight Security. *DNSDB*. URL: https://www.dnsdb.info.

[27] 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.

[28] 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.

[29] 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: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security - ASIA CCS ’17*. ACM Press, 2017, pp. 575–586.

[30] 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.

[31] Lihua Yao. “Reporting Valid and Reliable Overall Scores and Domain Scores”. In: *Journal of Educational Measurement* 47.3 (2016), pp. 339–360.

[32] 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.

[33] W.R. Gilks, S. Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in Practice*. 1st ed. Chapman & Hall, 1996.

[34] *mc-stan*. URL: http://mc-stan.org/.

[35] Andrew Gelman, Donald B Rubin, Andrew Gelman, and Donald B Rubin. “Inference from Iterative Simulation Using Multiple Sequences Linked references are available on JSTOR for this article : Inference from Iterative Simulation Using Multiple Sequences”. In: *Statistical Science* 7.4 (1992), pp. 457–472.

[36] Benjamin Edwards, Jay Jacobs, and Stephanie Forrest. “Risky Business: Assessing Security with External Measurements”. 2016.

[37] 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.

[38] 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.

[39] 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: *In Proceedings of the Research in Attacks, Intrusions and Defense (RAID)* (2016), pp. 1–20.

[40] 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.