# Apples, Oranges and Hosting Providers: Heterogeneity and Security in the Hosting Market

Samaneh Tajalizadehkhoob, Maciej Korczyński, Arman Noroozian, Carlos Gañán and Michel van Eeten
Faculty of Technology, Policy and Management, Delft University of Technology,
Delft, the Netherlands

***Abstract*— Hosting services are associated with various security threats, yet the market has barely been studied empirically. Most security research has relied on routing data and equates providers with Autonomous Systems, ignoring the complexity and heterogeneity of the market. To overcome these limitations, we combined passive DNS data with WHOIS data to identify providers and some of their properties. We found 45,434 hosting providers, spread around a median address space size of 1,517 IP addresses. There is surprisingly little consolidation in the market, even though its services seem amenable to economies of scale. We applied cluster analysis on several measurable characteristics of providers. This uncovered a diverse set of business profiles and an indication of what fraction of the market fits each profile. The profiles are associated with significant differences in security performance, as measured by the uptime of phishing sites. This suggests the approach provides an effective way for security researchers to take the heterogeneity of the market into account.**

## I. Introduction

Hosting providers play a pivotal role in the provisioning of all kinds of Internet-based services, as well as in mitigating the abuse of these services. Criminals purchase or hack services for hosting malware, phishing pages, command and control (C&C) servers, drop zones, dark markets, child pornography and more.

Over the years, various policies, standards and practices have emerged to improve hosting security (e.g., [1], [2]). These initiatives run into a significant barrier: the incredible complexity and heterogeneity of the hosting market. Even the most basic facts are unknown: How many providers are there? What address space do they manage? How are they distributed in terms of geography, size, types of services?

Developing policies and best practices in the absence of this kind of information seems unlikely to be effective. We cannot generate reliable security metrics for hosting providers without accounting for their heterogeneity [3]. It makes a big difference whether a best practice is geared towards hosting behemoths like GoDaddy, which operates an infrastructure across 800,000 IP addresses, towards the tiny providers which administer services on a single IP address, or perhaps towards some median point on this scale.

The authors thank Paul Vixie and Eric Ziegast from Farsight Security for sharing DNSDB and Thorsten Kraft from Cyscon for providing up-time data on phishing websites. This work was supported by NWO (grant nr. 12.003/628.001.003), the National Cyber Security Center (NCSC) and SIDN, the .NL Registry.

978-1-5090-0223-8/16/$31.00 

By necessity, security practices will look different across this spectrum. One can speculate that the same holds for security performance. Tiny providers might not be able to achieve the same level of competency the large providers with their dedicated abuse departments, but perhaps they make up for it by being more agile.

Remarkably, the complexity of the hosting market has barely been studied empirically, least of all in the area of security. Research has typically equated providers with Autonomous Systems [4]–[6]. Using routing data to identify providers and attribute security incidents is problematic as a lot of address space that is announced by an AS is not actually assigned to, or administered by, the AS owner.

There are some proprietary approaches to more accurately map the hosting space [7], but the underlying methodology and data are not publicly available. Lists published by sites like `webhosting.info` are of poor quality and lack key properties needed for research. In short: a decent map of the landscape is missing.

In this paper, we propose a novel measurement approach for capturing the complexity of the hosting market. In Section II, we systematically identify hosting providers through a fine-grained method combining passive DNS data to find hosting infrastructure and WHOIS data to determine address space assignment around that infrastructure. This results in a set of 45,434 hosting providers. Section III discusses the hosting landscape by exploring different provider characteristics that can be extracted from the data. In Section IV, we condense the complexity and heterogeneity of the hosting market by performing cluster analysis on the properties of providers. Finally, we demonstrate the value of these clusters by showing that they are associated with significant differences in the uptimes of phishing sites.

As far as we know, this is the first comprehensive mapping of the hosting provider market. The value of a more accurate mapping of the hosting market consists of 1) identification of providers rather than owners of Autonomous Systems; 2) more accurate attribution of security incidents to providers; 3) more accurate comparison and benchmarking of providers, also by normalizing for the size of providers. In the paper, we demonstrate these contributions by using the new map in a case study of phishing websites. We make the map available to other researchers upon request.

## II. Methodology for Identifying Hosting Providers

Message, Mobile and Malware Anti-Abuse Working Group (M3AAWG), a leading industry association, defines a hosting provider as "any entity which offers end users the ability to create their web presence on hardware they do not actually own" [1]. Hosting providers offer a variety of hosting services. They range from free and shared hosting services with limited resources and administrative privileges for customers, to more expensive services such as dedicated hosting and virtual private servers (VPS) where customers have more control over the computing resources [1]. The role of the provider to safeguard security also changes across these services.

While web presence is just one of the services on offer, we assume that all hosting providers have at least some webhosting in their portfolio. This allows us to use domain names as a way to identify providers. More specifically, we follow several steps to get from domain names to the population of providers (see Figure 1):

1) Extract domain names from DNSDB, a passive DNS dataset with an reasonable approximation of all domains in use on the web;
2) Identify the IP addresses where these websites are hosted;
3) Extract from WHOIS the netblocks to which these IP addresses belong and the organizations to which they are assigned;
4) Filter out the organizations that are clearly not hosting providers.

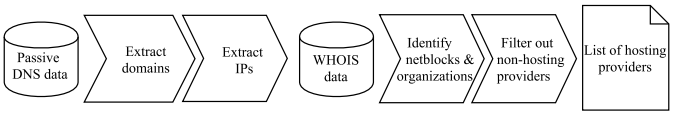


Fig. 1. Steps towards identifying hosting providers

In the next subsection, we systematically walk the reader through the design decisions taken in each step of this process.

### A. Identifying webhosting infrastructure

We first obtain a list that approximates the population of all domains in DNSDB – a passive DNS database that is generously shared with us by Farsight Security. To our knowledge, DNSDB has the best coverage of the overall domain name space that is available to researchers. It draws on hundreds of sensors worldwide and on the authoritative DNS data that various top-level domain (TLD) zone operators publish [8].

From DNSDB we extract all second-level domain names seen between January-June 2015 and the IP addresses that they resolved to. We identify 214,138,467 unique 2nd-level domain names that are mapped to 47,446,082 unique IP addresses.

### B. Identifying organizations and IP ranges

We use WHOIS data provided by Regional Internet Registries (RIR) to map the IP addresses of domains to the netblocks and names of the organizations to which these addresses are assigned.

WHOIS data has its own limitations, most notably the fact that records can be stale, inaccurate and non-standardized [9]. That being said, compared to the routing (BGP) data that most security research uses to associate IP addresses with providers, IP assignment better captures who is responsible for an address range and the services offered there than AS-level routing information. An AS, think of a data center, can announce routes for many different providers using its infrastructure.

We have used MaxMind's Organization database [10], which collates the WHOIS data of RIRs. The organization is identified by MaxMind from different fields of WHOIS databases, such as "descr" or "role" or "organization", depending on the RIR's WHOIS format.

When mapping IP addresses to organization names, an organization might appear multiple times in slightly different versions: `Go Daddy Netherlands B.V.`, `GoDaddy.com, LLC` and `GoDaddy.com Singapore`. The different names may point to the same organization. Sometimes, however, the differences reflect the fact that there are separate entities, for example in different jurisdictions. There currently is no reliable process to distinguish these situations, which is why we chose to not merge organizations with similar names.

Mapping IP addresses to their ranges and organizations results in a list of 161,891 organizations, covering 28,489 unique ASNs. On average, an ASN has address space allocated to around 7 organizations. This underlines just how problematic the current practice is to equate ASes with providers.

### C. Filtering out non-hosting providers

Clearly, not all of the organizations that host domains are hosting providers. When filtering out these cases, one has to balance potentially removing true positives versus keeping in false positives. Since our aim is to capture the complexity of the market, we do not want to lose true positives and apply three filters that conservatively remove false positives.

**Filter 1: AS level.** In a previous study [11], we have manually categorized 2000 ASes that contributed the most machines to botnet populations seen in sinkholes and spamtraps. Based on different data sources, we assigned ASes to one of the following types: **(*i*)** education, **(*ii*)** government, **(*iii*)** hosting, **(*iv*)** ISP-mobile, **(*v*)** ISP-other, **(*vi*)** ISP-broadband, and **(*vii*)** corporate networks such as banks, hospitals, etc. The first filter removes 6598 organizations (4% of the total set) that are located in the 332 ASes belonging to the categories education, government, and corporate networks.

**Filter 2: Organization level.** We generated a list of keywords for education, government and corporate networks. For example, the education category consist of the following list of keywords: universi, institut, college, school, akademi, academy, academi, research, teach, education, and science. We matched the keywords with organization names. In case of a match, we excluded the organization.

In this step we removed 39,369 organizations from the 155,293 that remained after the previous filter (25,4%), most of which matched an education keyword.

**Filter 3: Number of domains.** The third filter looks at the number of domains hosted by the organization. Organizations that host fewer domains than a certain threshold value are considered as "non-hosting". We hypothesize such organizations are not providing hosting services for others but instead they host their own websites.

To find the appropriate threshold, we took a sample of 163 organizations through a stratified sampling method to maintain the population's distribution in terms of the size of their address space, while keeping the sample size amenable to manual inspection. We manually assign "hosting" and "non-hosting" labels to the organizations by checking their names and visiting the corresponding websites, if they exist. The "hosting" label is assigned to all organizations that offer hosting service as a part of their business.

We then perform a sensitivity analysis on the threshold value for the number of domains to filter out non-hosting providers from the total set. For each threshold on the number of domains, we calculate the following parameters:

$$FP\ rate = \frac{FP}{FP+TN}, \qquad TP\ rate = \frac{TP}{TP+FN} \tag{1}$$

$$Accuracy = \frac{TP+TN}{FP+TN+TP+FN} \tag{2}$$

Where true positive (TP) is when an organization is correctly classified as "hosting", false positive (FP) is when an organization with "non-hosting" label is incorrectly classified as a hosting provider. Similarly, true negative (TN) is when an organization that is labeled as "non-hosting" is correctly classified as "non-hosting", whereas false negative (FN) is when an organization that has "hosting" label is incorrectly classified as "non-hosting".

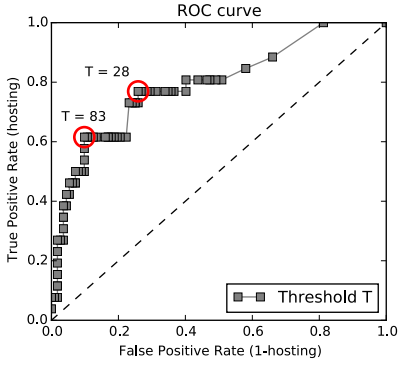


Fig. 2. ROC curve of different threshold values for the number of domains.

The receiver operating characteristic (ROC) curve shows the performance of different thresholds (Figure 2). The two thresholds marked with red circles in the ROC curve (T=83 and T=28) are the optimal thresholds for detecting hosting providers according to Equation 2. Note that our data is highly skewed and contains a large number of organizations with only a few domains. This leads to substantial noise when detecting hosting providers. At both thresholds, we have already included more than 99% of the total domain space in the data. Therefore the choice is essentially driven by the conservative approach of maximizing the chance of correctly identifying hosting providers and we are less sensitive to include false positives. We select T=28 as the threshold and define a hosting provider an organization that is hosting more than 28 domains. The filter discards 73,801 organizations from the set of 119,235 providers (62%) – e.g., Family Dental of Chicago (netblock 72.54.46.208/29) and United States Institute of Peace (netblock 64.210.233.0/23).

After applying these filters, we have a population of 45,434 organizations identified as hosting providers.

## III. Exploring the Hosting Landscape

From the underlying data, we can extract several characteristics of the 45,434 hosting providers, such as the size of their address space, as well as the portion of that space used for webhosting. What can these tell us about the hosting market?

**[c1] IP address range size**: The first plot in Figure 3 displays the distribution of providers in terms of their address space. The distribution goes from around 200 providers with only one IP addresses all the way up to providers with six or seven orders of magnitude larger address space. There we find ISPs like AT&T and Comcast, for whom hosting is not the main service. The distribution is centered around providers with 1,000 to 10,000 addresses (median: 1,517). From an economic perspective, this market shows a surprising lack of consolidation. One would think that economies of scale, in combination with commoditized services that can be globally delivered, would lead to a few large providers dominating the market. This mechanism is clearly visible in cloud services, but not here. It takes 1,210 providers to account for 80% of the address space used for webhosting. How can the many medium-sized providers compete on price with the large ones? How do the tiny providers survive in this market? This finding underlines that we know little about the incentives in this market and the security practices that they give rise to.

**[c2] Percentage of IP range used for hosting websites**: What percentage of the address space of a provider is used for webhosting? This tells us to what extent webhosting is the core business model or not. The second plot shows the distribution of providers. It shows that for the bulk of them, webhosting is only a minor part of their infrastructure. Their infrastructure might be used to run game servers, databases, VPN exit nodes, and other services. Some smaller providers use almost all of their address space for webhosting, whereas larger companies such as GoDaddy and OVH are using approximately half of their allocated range for webhosting, but they are all on the higher end of the spectrum.

**[c3] Percentage of IP range used for shared hosting**: When talking about abuse in hosting services, shared hosting is often flagged as a problem area [12], [13]. One reason is the low profit margins of these services, which seems to be accompanied by poor security, according to a recent study [14]. The third plot shows the percentage of the address space

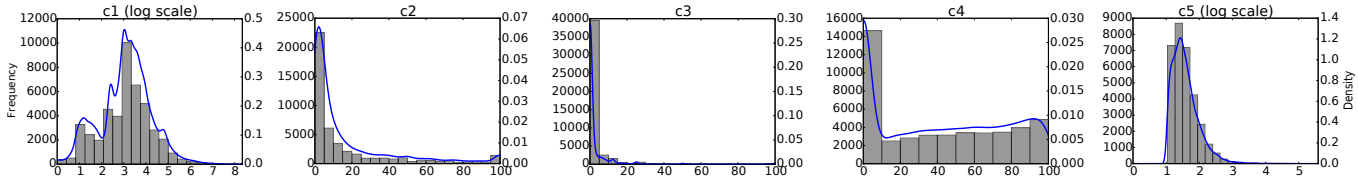


Fig. 3. Histograms and kernel density estimates for five characteristics of hosting providers

used for shared hosting. We consider an IP address to be used for shared hosting if it serves more than 10 domains. While shared hosting draws a lot of attention in research, most providers actually use only about 10% of their address space for this purpose. Only around 500 providers use more than 50% for shared hosting, while 225 focus exclusively on shared hosting.

**[c4] Percentage of domain names on shared hosting**: A slightly different take on the importance of shared hosting is to look at its portion of all domains that are hosted by the provider. The fourth plot shows a rather uniform distribution, except for the first group, who offer no shared hosting at all. In other words, for webhosting as a service, shared hosting is provided in all portfolios and has a fluid proportion to other webhosting solutions, like VPS or dedicated hosting.

**[c5] Density of domains on shared hosting IP addresses**: The average number of domain names on IP addresses used for shared hosting can indicate in what part of the market the provider is competing. Higher density (more domains per server) would indicate more shared resources and competing for lower value customers. The last plot shows that a few hundreds of providers have shared IP addresses with more than thousand domains, on average, while the majority of the providers have 10 to 100 domains per shared IP address.

These individual characteristics give us a sense of the hosting landscape. We can see just how much complexity and heterogeneity is present across providers. There is remarkably little consolidation and many small players shape the landscape as much as the larger providers. Webhosting, the services that has dominated the image of the sector, only plays a limited role for many providers – and shared hosting even more so.

All of these characteristics influence security incentives and practices, especially in combination. Viewing a characteristic isolated from the others can be misleading. For example, when looking at the influence of size of a provider, one cannot simply use address space as a proxy, because it ignores the fact that the providers with the largest address space are not predominantly hosting providers, so their hosting product groups may actually resemble those of small or medium-sized providers.

To deepen our understanding of the market, we would need to identify how different values for these characteristics occur in combination across the population of providers. We propose to profile the providers by performing cluster analysis on the characteristics. This would condense the complexity into a tractable starting point for further empirical research. Are certain types of providers more effective in securing their infrastructure? Perhaps type is not that relevant. We might find a equally strong and poor security practices within each type. In the remainder of this paper we first perform cluster analysis on the characteristics and then use those clusters to determine whether they uncover meaningful differences in terms of security, as measured by the uptime of phishing websites in the networks of these providers.

## IV. Categorizing Hosting Providers

We try to profile hosting providers using the set of five characteristics explained in the previous section. We first identify the appropriate algorithm and then discuss the clusters.

### A. Choice of the clustering algorithm

To meaningfully partition the hosting space, we test four clustering algorithms: *k*-means, *k*-medoids, expectation maximization (EM), and hierarchical. We first randomly sample ten thousands hosting providers, we then evaluate the four selected algorithms using five types of cluster validation measures, as described by Brock *et al.* [15]. Table I reports on the stability metrics (APN: average proportion of non-overlap, ADM: average distance between means, and FM: figure of merit) and internal metrics (connectivity and silhouette width) calculated for four clustering algorithms and different numbers of clusters.

The results shown in Table I indicate that clustering of hosting providers obtained using hierarchical and *k*-means algorithms are more stable (smaller values of APN and ADM) and compact (lower connectivity and Silhouette width close to 1) comparing to *k*-medoids and EM algorithms. Given the similarity in evaluation results of *k*-means and hierarchical algorithms, we choose the former. It is computationally more efficient and it enables the iterative improvements in grouping of the hosting providers. We inspected the stability and internal metrics as a function of a number of clusters (see Table I). Combined with our domain knowledge about the hosting sector, we grouped providers in 10 clusters using *k*-means.

### B. Groups of hosting providers

Table II shows the groups: the size (number of providers), and the mean and standard deviation of each characteristic. Cluster 2 represents a group of the smallest hosting providers that are assigned on average a few to a few dozen IP addresses which are only used for shared hosting–the proportion of provider's domain name space and IP space used for shared hosting ($c3$ and $c4$) is above 97%. Note that the mean density of domains per shared hosting IP address ($c5$) is very high (1720), as is the standard deviation. Both are driven by one provider with an astonishing 385,757 domains registered to a single IP address (other sources, like `DomainTools.com` report this as well). We have no information on what business

TABLE I
STABILITY AND INTERNAL METRICS PER CLUSTERING ALGORITHM AND NUMBER OF CLUSTERS

| Clustering Algorithm | Metric | Number of Clusters | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| hierarchical [16] | APN | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | ADM | 10,628.70 | 17,804.84 | 17,805.16 | 17,805.16 | 21,989.46 | 21,991.56 | 21,993.01 | 21,993.08 | 26,749.91 | 26,750.06 | 26,751.00 | 26,751.00 | 29,675.66 |
| | FOM | 422,786.09 | 422,807.23 | 422,828.35 | 422,849.48 | 422,870.63 | 422,891.67 | 422,912.59 | 422,933.65 | 422,954.74 | 422,975.74 | 422,996.68 | 423,017.85 | 423,039.01 |
| | Connectivity | 3.86 | 9.54 | 11.54 | 14.31 | 18.87 | 20.87 | 23.37 | 27.10 | 29.42 | 31.42 | 35.47 | 40.32 | 42.92 |
| | Silhouette | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.98 |
| kmeans [17] | APN | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.02 |
| | ADM | 10,624.97 | 18,525.27 | 18,527.03 | 21,989.46 | 25,463.27 | 28,099.38 | 28,100.34 | 28,289.60 | 28,290.13 | 29,999.74 | 30,000.07 | 30,085.53 | 30,112.69 |
| | FOM | 422,786.17 | 422,807.20 | 422,828.28 | 422,849.48 | 422,870.63 | 422,891.68 | 422,912.71 | 422,933.64 | 422,954.50 | 422,975.51 | 422,996.68 | 422,904.85 | 422,925.75 |
| | Connectivity | 3.86 | 11.53 | 13.53 | 16.82 | 18.23 | 22.09 | 24.09 | 29.52 | 32.02 | 32.74 | 34.74 | 46.86 | 45.58 |
| | Silhouette | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.98 | 0.98 | 0.98 | 0.98 |
| kmedoids [18] | APN | 0.00 | 0.10 | 0.11 | 0.14 | 0.14 | 0.14 | 0.15 | 0.15 | 0.16 | 0.16 | 0.16 | 0.16 | 0.16 |
| | ADM | 10,622.25 | 25,458.85 | 27,179.76 | 30,109.91 | 31,577.42 | 31,581.43 | 31,350.52 | 31,580.66 | 31,662.98 | 31,790.83 | 31,663.69 | 31,668.07 | 31,708.57 |
| | FOM | 422,786.18 | 422,806.49 | 422,686.58 | 422,708.06 | 422,751.59 | 422,768.74 | 422,661.05 | 422,690.55 | 422,696.29 | 422,717.51 | 422,632.32 | 422,653.60 | 422,662.08 |
| | Connectivity | 3.86 | 5.38 | 13.72 | 16.33 | 18.47 | 29.76 | 31.76 | 44.55 | 47.33 | 39.30 | 45.86 | 57.83 | 60.45 |
| | Silhouette | 1.00 | 0.99 | 0.99 | 0.98 | 0.87 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.73 | 0.73 | 0.74 |
| EM [19] | APN | 0.18 | 0.27 | 0.49 | 0.38 | 0.46 | 0.47 | 0.52 | 0.49 | 0.48 | 0.50 | 0.53 | 0.48 | 0.51 |
| | ADM | 85,712.28 | 101,145.64 | 104,900.13 | 118,009.14 | 116,255.66 | 114,734.28 | 101,111.36 | 117,688.39 | 99,655.42 | 120,131.56 | 112,749.72 | 100,706.27 | 109,422.14 |
| | FOM | 422,785.45 | 422,788.48 | 422,793.21 | 422,807.16 | 422,790.86 | 422,860.83 | 422,604.51 | 422,635.49 | 422,573.50 | 422,601.98 | 422,525.90 | 422,590.41 | 422,520.98 |
| | Connectivity | 1,400.55 | 2,589.21 | 3,022.55 | 4,338.44 | 4,692.36 | 5,689.08 | 6,378.72 | 7,345.93 | 7,572.52 | 7,349.92 | 7,742.04 | 7,034.40 | 8,122.14 |
| | Silhouette | 0.39 | -0.10 | 0.11 | -0.33 | -0.52 | -0.52 | -0.51 | -0.51 | -0.51 | -0.46 | -0.49 | -0.45 | -0.45 |

model is at work here. Without this provider, the mean density drops to 178 (SD: 222). Providers in this cluster are mainly located in United States (97.2%). They offer a great variety of cheap or even free hosting services. For example, we observed an average of 1983 domains per IP hosted on 2048 IP addresses of the OpenTLD Web Network TK organization (the `.tk` registry). Most of these providers include free plans with limited web space and data transfer under certain second-level domains. For a monthly fee of few euros a customer may obtain an unlimited number of domains under the most popular gTLDs as well as unlimited storage and bandwidth.

Clusters 4 and 8 contain somewhat larger providers (in terms of address space) such as 1&1 Internet. They offer more diverse services in comparison to the smaller ones. Around 80% of their addresses are used for webhosting, but only a small share of this space is used for shared hosting services (11% and 33%, respectively). The lower density of domains over shared IP addresses ($c5$) may suggest that they offer virtual private hosting as an extension for the hosting services. This type of service is usually unmanaged, i.e., the customer administrates the virtual system and software that runs on the server.

Cluster 10 is similar to clusters 4 and 8 in terms of the mean size of the IP range ($c1$) and the portion of IP space used for webhosting ($c2$), while a much smaller portion of the address space ($c3$) and domain name space ($c4$) is shared hosting. This suggest that providers in this cluster offer more non-shared (and thus expensive) type of services, such as dedicated hosting.

Clusters 3 and 7 are the next class in terms of size, moving from hundreds to thousands of IP addresses ($c1$). A smaller portion of the address space is for webhosting ($c2$)—40% and 33% respectively. The providers in cluster 3 use 7.6% of their IP address and 80% of their domain name space for shared hosting ($c3$ and $c4$). In cluster 7, on the other hand, providers have less shared hosting address space (0.49%) and only 9.85% of all domains are on shared addresses. Again, this suggests that providers in this group such as Go Daddy offer more dedicated or managed hosting services.

TABLE II
HOSTING PROVIDER GROUPS

| Cluster | Size | Mean (Standard Deviation) | | | | |
|---|---|---|---|---|---|---|
| | | c1 | c2 | c3 | c4 | c5 |
| 1 | 7,413<br>16,81% | 48286.64<br>(263086.31) | 4.36<br>(4.90) | 0.15<br>(0.25) | 31.20<br>(8.16) | 25.97<br>(19.41) |
| 2 | 250<br>0.47% | 28.44<br>(205.85) | 99.62<br>(2.46) | 97.77<br>(7.43) | 99.68<br>(1.79) | 1720.09<br>(24387.17) |
| 3 | 3,771<br>7.94% | 2441.18<br>(29297.19) | 40.03<br>(10.69) | 7.58<br>(4.51) | 80.64<br>(15.00) | 114.57<br>(402.76) |
| 4 | 1,748<br>3.35% | 210.35<br>(2455.62) | 81.00<br>(13.86) | 10.92<br>(4.75) | 75.93<br>(15.72) | 117.74<br>(1354.02) |
| 5 | 13,367<br>29.01% | 48775.60<br>(377535.49) | 4.77<br>(4.90) | 0.01<br>(0.03) | 1.96<br>(4.38) | 4.08<br>(10.67) |
| 6 | 6,657<br>15.02% | 16594.20<br>(101181.78) | 6.83<br>(6.62) | 1.31<br>(2.26) | 85.43<br>(8.33) | 391.45<br>(3946.55) |
| 7 | 2,550<br>5.90% | 5948.00<br>(75045.93) | 33.08<br>(9.98) | 0.49<br>(1.01) | 9.85<br>(14.34) | 11.30<br>(22.71) |
| 8 | 988<br>1.88% | 459.95<br>(9557.63) | 79.14<br>(21.34) | 32.82<br>(10.71) | 91.78<br>(9.17) | 113.18<br>(1006.72) |
| 9 | 7,389<br>17.07% | 307011.67<br>(4327995.09) | 4.95<br>(5.75) | 0.35<br>(0.54) | 57.85<br>(8.40) | 42.40<br>(46.62) |
| 10 | 1,301<br>2.55% | 679.87<br>(6270.40) | 79.21<br>(15.22) | 0.98<br>(1.98) | 8.13<br>(13.89) | 8.10<br>(20.77) |

*c1: IP address range size*
*c2: Percentage of IP address range used for hosting websites*
*c3: Percentage of IP address range used for shared webhosting*
*c4: Percentage of domain names on shared webhosting*
*c5: Density of domains on shared webhosting IP addresses*

Similar conclusions could be drawn from a comparison of hosting providers in clusters 5 and 6. We move, once more, up one class in terms of size of the address space, where the portion of those addresses used for webhosting further diminishes. In contrast to cluster 5, the webhosting of providers in cluster 6 is mostly shared hosting. In comparison to other clusters, cluster 5 has the smallest shared hosting portion of its IP address space and domain name space.

Cluster 1 is similar to cluster 5 in terms of the allocated IP space ($c1$) and the portion of IP space used for webhosting

($c2$) while a bigger portion of domain name space ($c4$) in this cluster is shared hosting.

Cluster 9 with around 17% of the total providers in the data, mostly contains providers with the largest allocated IP space ($c1$) and a small portion of the address space used for webhosting ($c2$) such as Telecom companies. The values of percentage of IP space and domain names space used for shared hosting ($c3$ and $c4$ respectively) suggest a significant portion of the webhosting in this group is shared hosting.

These results, while crude, allow us to distinguish groups of providers with different profiles, from small companies that offer cheap webhosting on highly dense shared servers from those providers that offer more expensive and flexible services, such as managed and dedicated hosting.

Finally, we analyze the geographical location of the providers in each of the clusters. Most of the providers in clusters with smaller average IP ranges are located in United States while clusters containing providers with larger IP range sizes are evenly distributed across different countries.

We expect that different groups of providers offering various types of hosting services handle domain abuse differently, which is then examined in terms of uptimes of phishing domains discussed in Section V.

## V. Case Study: Analysis of Uptime for Phishing Websites

In the previous sections, we grouped the hosting providers into 10 different clusters with different business profiles. In this section, we examine whether these profiles are associated with differences in abuse handling, more specifically, the speed with which phishing websites are taken down.

### A. Phishing data

We analyze data on the uptime of phishing websites from the moment the provider has been notified, which was generously provided to us by Cyscon GmbH [20].

TABLE III
Summary of phishing data points per cluster

| Cluster | Providers | ASes | FQDNs | URLs | IPs | Countries |
|---|---|---|---|---|---|---|
| 1 | 221 | 229 | 633 | 3234 | 367 | 63 |
| 2 | 24 | 6 | 425 | 556 | 241 | 4 |
| 3 | 453 | 357 | 29641 | 78592 | 8036 | 54 |
| 4 | 86 | 41 | 689 | 1418 | 344 | 19 |
| 5 | 82 | 84 | 210 | 2521 | 134 | 43 |
| 6 | 938 | 893 | 10265 | 30638 | 4998 | 84 |
| 7 | 47 | 48 | 465 | 1400 | 229 | 21 |
| 8 | 48 | 19 | 1130 | 1634 | 734 | 11 |
| 9 | 483 | 504 | 4165 | 13957 | 1677 | 77 |
| 10 | 12 | 12 | 155 | 482 | 98 | 6 |

The dataset contains 137,577 phishing URLs associated with 48,224 fully qualified domain names (FQDNs) that were hosted on 17,279 IP addresses in 1,962 ASes located in 114 countries. Each websites is then tagged with the first and last time it is seen online. Note that for the websites that are only seen once, the first seen is the same as the last seen, indicating that they were taken down before the second measurement moment. These are logged as having an uptime of 0 hours. The data contains websites that were first seen between June 4 to August 16, 2015. Many of the targets are known brands such as Paypal, Dropbox, Yahoo, or Wells Fargo, World of Warcraft and Battlenet.

We mapped the phishing data to the different clusters of hosting providers discussed in Section IV. Table III displays the distribution of the data across the different clusters.

### B. Analysis of uptime

An important criteria to evaluate security performance of hosting providers, is how fast they respond to being notified about malicious sites [4]. Uptime has been used in previous security research as a standard metric for studying lifetime of different attack types [21]–[23].

We define "uptime" of a phishing website as the number of days between the first and last time the phishing site is observed online and reported by Cyscon. Some of the phishing sites remain online beyond the measurement period, which leaves their uptime unknown. To correctly account for these cases, we analyze uptimes through survival analysis with right-censoring.

The survival function $S(t)$ expresses the probability that a phishing website is online at a specific time during the observation period. It is calculated at time $t$ using the standard Kaplan-Meier estimator without any assumption about the distribution of the underlying data [24].

Figure 4 shows survival curves for phishing websites in the different provider clusters. In Table IV we present descriptive statistics on uptimes, based only on sites that had been taken down by the end of our measurement period.

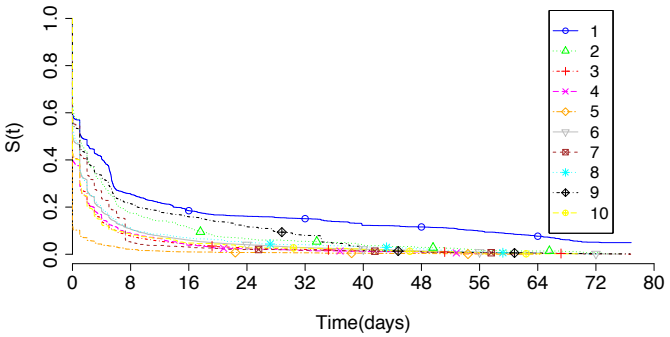


Fig. 4. Kaplan-Meier estimates per cluster

TABLE IV
Descriptive statistics of uptimes (hours) per cluster

| Cluster | Min | Mean | Median | Max | SD | Coef Var | SE |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 165.840 | 24.000 | 1,744.400 | 344.390 | 207.660 | 6.315 |
| 2 | 0 | 142.360 | 24.294 | 1,813.800 | 301.530 | 211.810 | 12.393 |
| 3 | 0 | 59.408 | 0.0003 | 1,829.900 | 175.280 | 295.050 | 0.627 |
| 4 | 0 | 62.560 | 0.0003 | 1,505.200 | 175.220 | 280.080 | 4.650 |
| 5 | 0 | 16.715 | 0 | 1,542.600 | 98.499 | 589.290 | 1.960 |
| 6 | 0 | 80.498 | 0.002 | 1,812.800 | 210.080 | 260.980 | 1.176 |
| 7 | 0 | 76.794 | 24.004 | 1,723.100 | 182.630 | 237.820 | 4.876 |
| 8 | 0 | 95.504 | 5.005 | 1,730.100 | 249.650 | 261.400 | 6.168 |
| 9 | 0 | 152.790 | 24 | 1,840.800 | 276.420 | 180.910 | 2.307 |
| 10 | 0 | 70.064 | 0.0003 | 1,671.600 | 205.000 | 292.580 | 9.318 |

The differences among the survival curves are highly significant, not only across the population as a whole, but even

when performing pair-wise comparisons among all clusters. Figure 5 displays the results for log-rank non-parametric tests [25]. Only the blank tiles indicate non-significant differences at a 0.05 significance level. In other words, the different clusters are associated with different security performance. This underlines the value of the preceding work of mapping and then condensing the complexity and heterogeneity of the hosting market. Explaining the differences in uptimes from the properties of the providers in the clusters is beyond the scope of this paper and the topic of our ongoing work. We can, however, explore what these results show, without drawing any hard conclusions.

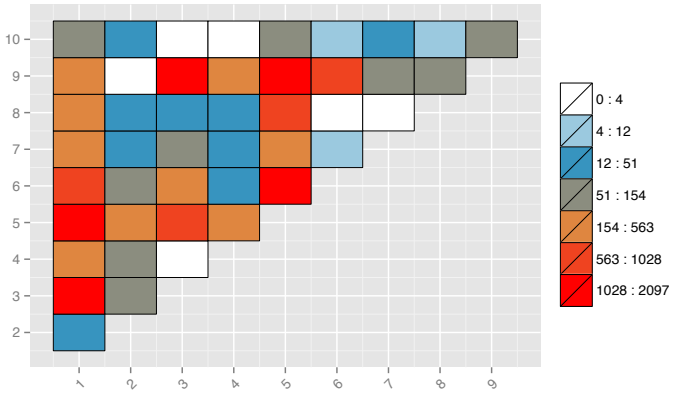


Fig. 5. Log-rank test for cluster pairs

Figure 4 and Table IV shows that phishing websites in clusters 1 and 9 have the highest survival rate – in other words, these clusters perform the worst in terms of take-down speed. Clusters 1 and 9 contain the largest providers in the market (together with cluster 5, see Table II). Providers in these two clusters have a relatively low percentage of webhosting, but a significant portion of that webhosting is shared hosting. Around half of the phishing sites in these clusters indeed map to shared hosting servers. The third-worst performer is cluster 2. This contains the providers with the smallest allocated IP ranges, which are used completely for shared hosting services.

An intriguing contrast emerges when looking at the best performer: cluster 5. It is very similar to 1 and 9, except for the fact that it contains virtually no shared hosting. It is too early to draw any conclusions from these findings, but it seems clear that size of the allocated address space itself does not explain performance. Perhaps it is more related to the position and size of shared hosting services in the overall portfolio. This is consistent with earlier security research that focused on shared hosting as a problem area. The underlying economic mechanism would be that this part of the market is driven by fierce price competition and low profit margins.

Whether the uptimes of phishing sites are really related to the incentives and practices around shared hosting is a question for future work. In a more general sense, our findings demonstrate that a better mapping of the market and its providers will allow us to focus security efforts in the most urgent areas, as well as allowing us to compare apples to apples when evaluating the security of different providers.

## VI. Related Work

To the best of our knowledge, there is no study that has systematically and transparently mapped the hosting market. Recent work by Noroozian *et al.* underlines the need for such mapping, by demonstrating how provider heterogeneity influences security performance metrics [3].

A number of studies map security incidents to hosting providers by equating them with ASes and normalizing the incidents by the AS size [5]. Mahjoub studies the concentration of maliciousness in ASes by analyzing AS topology, hosted content and IP space reservation [26]. Other studies identify malicious ASes using AS topology, BGP-related features and by exploring ASes providing transit for malicious ASes [6], [27]. Although useful, these studies neglect the organizations within ASes and their properties, which influence all metrics of maliciousness.

Industry is more active in producing rankings for ASes as hosting providers– e.g., [28]. Netcraft's uses reverse DNS to map providers, but the complete methodology and data are not available to researchers [7]. Canali *et al.* examine the security performance of a small group of shared hosting providers and conclude that the majority of the providers are unable to detect even basic attacks on their networks [14]. Although they study providers with specific characteristics, the sample of providers is non-random and too small to draw any conclusions about providers in general.

A separate branch of research focuses more on how hosting providers deal with the take-down of malicious websites [4], [29]. Nappa *et al.* explore lifetime of drive-by download URLs and rank their associated ASes [30]. Moore and Clayton study lifetime of phishing domains and variables like hosting providers of the website that might influence take-down speed and conclude that website removal is not yet fast enough to completely mitigate the problem of phishing [22]. Gañán *et al.* examine characteristics of botnet C&Cs that might influence their lifetime [31]. Again, treating providers as ASes, the paper concludes that hosting provider, hosting types (e.g., bulletproof or free) and popularity of the sites are significant factors associated with the uptime of the C&Cs.

We believe that this paper is the first to map the hosting market and discuss its heterogeneity by analyzing the differences among the providers in terms of their services and their abuse handling practices.

## VII. Conclusion and Discussion

A variety of initiatives seek to improve security in hosting services, but none of them have taken even basic information about the market into account, which makes it hard to identify best practices and evaluate performance. Security research has mostly relied on routing data and AS-level aggregations of security incidents, equating ASes to providers. To overcome these limitations, We have developed a systematic approach to uncover and grasp the complexity of the hosting market. We combined passive DNS data to determine the address space of hosting infrastructure with WHOIS data to determine the associated providers and their IP address space.

Next, we applied several filters to conservatively remove false positives (non-hosting providers). This process resulted in a set of 45,434 hosting providers, somewhat log-normally spread around a median size of 1,517 IP addresses. Using five provider characteristics we extracted from the data, we familiarized the reader with the hosting landscape. There is surprisingly little consolidation in the market, given that the services are commoditized and thus amenable to economies of scale, as can be seen in the market for cloud services. In hosting, it takes 1,210 providers to account for 80% of the address space used for webhosting. A large number of small players dominate the landscape as much as a small number of larger providers. There are providers with millions of IP addresses and around a thousand with a handful or even just a single address. We found providers who are offering only webhosting versus those who are using only a small share of their allocated address space for webhosting.

We explored what combinations of the characteristics occur in reality via cluster analysis. This uncovered a diverse set of business profiles and an indication of what fraction of the market fits each profile. Since these profiles are proxies for different types of organizations, we assessed whether the clusters were associated with different security performance using data on the uptime of phishing websites. The clusters were indeed very different in how fast they take down phishing domains. The results suggest that our mapping of the hosting market is helpful in deepening our understanding of the driving forces of security threats, as well as in developing best practices. Both benefit from being able to compare apples to apples, rather than using the current crude analytical approaches based on routing data and AS-level abuse metrics, which cannot account for the heterogeneity in the market.

Several limitations need to be acknowledged. These results are just a first step towards a thorough understanding of the market. We assumed that all providers offer at least some webhosting in their portfolio, so as to be able to use passive DNS data to identify potential providers. There might be some providers who do not offer webhosting. They would be invisible to this approach. Another limitation is the fact that WHOIS records are notorious for containing stale, inconsistent and inaccurate data. Related to this are the inconsistencies in organization names in the WHOIS data. When different names point to the same entity, they might actually be operated under one entity or they may point to entities belonging to the same parent company but operating independently of each other. How to distinguish these two cases is still unsolved. The accuracy of the filters to separate hosting providers from other entities that host websites is rather limited and this impacts the mapping.

Future work is needed to explain the significant differences that were found among the clusters of providers. The characteristics of providers can also be enriched by adding other variables that might shape their incentives and performance, such as their jurisdiction, privacy and security regulations and development indicators.

The map of the hosting provider landscape that has been developed in the course of this study will be made available to other researchers, so as to contribute to better analysis and mitigation of the security threats that plague this market.

## REFERENCES


[1] M3AAWG. (2015) M3AAWG anti-abuse best common practices for hosting and cloud service providers. [Online]. Available: https://www.m3aawg.org/sites/maawg/files/news/M3AAWG_Hosting_Abuse_BCPs-2015-03.pdf

[2] Stop Badware. (2011) Best practices for web hosting providers. [Online]. Available: https://www.stopbadware.org/files/best-practices-responding-to-badware-reports.pdf

[3] A. Noroozian, M. Korczyński, S. TajalizadehKhoob, and M. van Eeten, "Developing security reputation metrics for hosting providers," in *8th Workshop on Cyber Security Experimentation and Test (CSET 15)*. USENIX Association, 2015.

[4] B. Stone-Gross, C. Kruegel, K. Almeroth, A. Moser, and E. Kirda, "Fire: Finding rogue networks," in *Computer Security Applications Conference*. IEEE, 2009, pp. 231–240.

[5] C. A. Shue, A. J. Kalafut, and M. Gupta, "Abnormally malicious autonomous systems and their internet connectivity," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 220–230, 2012.

[6] C. Wagner, J. François, R. State, A. Dulaunoy, T. Engel, and G. Massen, "Asmatra: Ranking ASs providing transit service to malware hosters," in *IFIP/IEEE International Symposium on Integrated Network Management*. IEEE, 2013, pp. 260–268.

[7] Netcraft. (2015) Hosting provider server count. [Online]. Available: http://www.netcraft.com/internet-data-mining/hosting-provider-server-count/

[8] DNSDB. [Online]. Available: https://www.dnsdb.info/

[9] K. Elliott, "Who, what, where, when, and why of whois: Privacy and accuracy concerns of the whois database," *SMU Science & Technology Law Review*, vol. 12, p. 141, 2008.

[10] Maxmind GeoIP. [Online]. Available: http://dev.maxmind.com/geoip/legacy/downloadable/

[11] H. Asghari, M. van Eeten, and J. Bauer, "Economics of fighting botnets: Lessons from a decade of mitigation," *Security Privacy, IEEE*, vol. 13, no. 5, pp. 16–23, Sept 2015.

[12] T. Van Goethem, P. Chen, N. Nikiforakis, L. Desmet, and W. Joosen, "Large-scale security analysis of the web: Challenges and findings," in *Trust and Trustworthy Computing*. Springer, 2014, pp. 110–126.

[13] G. Aaron and R. Rasmussen. (2015) Global phishing survey Trends and domain name use in 1h2014. [Online]. Available: http://docs.apwg.org/reports/APWG_Global_Phishing_Report_1H_2014.pdf

[14] D. Canali, D. Balzarotti, and A. Francillon, "The role of web hosting providers in detecting compromised websites," in *Proceedings of the 22nd international conference on World Wide Web*. World Wide Web Conferences, 2013, pp. 177–188.

[15] G. Brock, V. Pihur, S. Datta, and S. Datta, "clValid, an R package for cluster validation," *Journal of Statistical Software*, 2011.

[16] F. Murtagh, *Multidimensional clustering algorithms*, 1985.

[17] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[18] L. Kaufman and P. J. Rousseeuw, *Partitioning Around Medoids (Program PAM)*. John Wiley & Sons, Inc., 2008, pp. 68–125.

[19] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. pp. 611–631, 2002.

[20] Cyscon GmbH. [Online]. Available: http://www.cyscon.de

[21] N. Leontiadis, T. Moore, and N. Christin, "Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade," in *USENIX Security Symposium*, 2011.

[22] T. Moore and R. Clayton, "Examining the impact of website take-down on phishing," in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*. ACM, 2007, pp. 1–13.

[23] J. Nazario and T. Holz, "As the net churns: Fast-flux botnet observations," in *International Conference on Malicious and Unwanted Software*. IEEE, 2008, pp. 24–31.

[24] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[25] R. Peto and J. Peto, "Asymptotically efficient rank invariant test procedures," *Journal of the Royal Statistical Society. Series A (General)*, pp. 185–207, 1972.

[26] D. Mahjoub, "Sweeping the IP space: The hunt for evil on the internet." Virus Bulletin Conference, 2014. [Online]. Available: https://www.virusbtn.com/pdf/conference/vb2014/VB2014-Mahjoub.pdf

[27] M. Konte, R. Perdisci, and N. 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*. ACM, 2015, pp. 625–638.

[28] Hostexploit. [Online]. Available: http://hostexploit.com/

[29] O. Cetin, M. H. Jhaveri, C. Ganán, M. van Eeten, and T. Moore, "Understanding the role of sender reputation in abuse reporting and cleanup," in *Workshop on the Economics of Information Security*. WEIS, 2015.

[30] A. Nappa, M. Z. Rafique, and J. Caballero, "Driving in the cloud: An analysis of drive-by download operations and abuse reporting," in *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2013, pp. 1–20.

[31] C. Gañán, O. Cetin, and M. van Eeten, "An empirical analysis of Zeus C&C lifetime," in *Proceedings of the 10th ACM Symposium (ASIA CCS)*. ACM, 2015, pp. 97–108.