Public Review for

The Web is Still Small After More Than a Decade

Nguyen Phong Hoang, Arian Akhavan Niaki, Michalis Polychronakis, and Phillipa Gill

Web topology has always been an important topic in measurement studies. Understanding the topological characteristics of the Internet, the web, and their growth and evolution has been instrumental to development of new protocols, content caching mechanisms, and detecting security issues.

In this paper the authors revisit some of the decade-old studies on web presence and co-location, and find that despite drastic changes in the way we use the Internet today, the web still maintains a largely centralised and core-centric DNS-based topology.

Although the papers' findings are not drastically surprising, the reviewers found that the paper and the resulting dataset are interesting and could make a useful contribution to the field and have recommended its publication.

Public review written by
Hamed Haddadi
Imperial College, UK

The Web is Still Small After More Than a Decade

A Revisit Study of Web Co-location

Nguyen Phong Hoang,[†][*] Arian Akhavan Niaki,[§][*] Michalis Polychronakis,[†] Phillipa Gill[§]
Stony Brook University, New York, USA[†] University of Massachusetts, Amherst, USA[§]
{nghoang, mikepo}@cs.stonybrook.edu {arian, phillipa}@cs.umass.edu

ABSTRACT

Understanding web co-location is essential for various reasons. For instance, it can help one to assess the collateral damage that denial-of-service attacks or IP-based blocking can cause to the availability of co-located web sites. However, it has been more than a decade since the first study was conducted in 2007. The Internet infrastructure has changed drastically since then, necessitating a renewed study to comprehend the nature of web co-location.

In this paper, we conduct an empirical study to revisit web co-location using datasets collected from active DNS measurements. Our results show that the web is still small and centralized to a handful of hosting providers. More specifically, we find that more than 60% of web sites are co-located with at least ten other web sites—a group comprising less popular web sites. In contrast, 17.5% of mostly popular web sites are served from their own servers.

Although a high degree of web co-location could make co-hosted sites vulnerable to DoS attacks, our findings show that it is an increasing trend to co-host many web sites and serve them from well-provisioned content delivery networks (CDN) of major providers that provide advanced DoS protection benefits. Regardless of the high degree of web co-location, our analyses of popular block lists indicate that IP-based blocking does not cause severe collateral damage as previously thought.

CCS CONCEPTS

• **Networks → Network measurement**;

KEYWORDS

Web co-location, DNS measurement, blocking collateral damage

1 INTRODUCTION

Over the last three decades, the World Wide Web (the web for brevity) has grown exponentially, thanks to the rapid expansion of the Internet. A web site is a fundamental unit that makes up the web, in which related web resources (e.g., web pages, images, audios, and videos) are gathered and published via a web server identified by a domain name (e.g., example.com). Prior to 1997, each web site was typically hosted on its own server with a distinct IP address. Therefore, the number of unique IP addresses, with the standard web port (i.e., 80) open, was an accurate proxy to estimate the number of web sites at the time. However, since the introduction of name-based virtual hosting technology as a part of HTTP/1.1 in 1997 [12], many web sites can be co-hosted on the same IP address, making it more challenging and sophisticated to measure the web, especially in terms of web co-location [25].

Understanding web site co-location is essential for various reasons. For instance, it can help one to assess the collateral damage that denial-of-service (DoS) attacks or IP-based blocking can cause to the availability of co-located web sites. Shue et al. [37] conducted the first study of web co-location more than a decade ago and found that the web was smaller than it seemed in terms of the location of servers. The study quantifies i) the extent to which the availability of the web can be affected by targeted DoS attacks, and ii) the impact of several IP block lists on co-hosted web sites. Since then, the Internet has grown dramatically. More than 354 million domain names have been registered across all top-level domains (TLDs) as of the second quarter of 2019 [41]. In addition, the adoption of IPv6 and CDNs have changed the way web traffic is delivered. Considering these drastic changes of the Internet infrastructure over the last decade, it is desirable to investigate whether previous findings by Shue et al. [37] still hold in today's web ecosystem.

In this paper, we revisit the study of web site co-location by analyzing datasets collected from active DNS measurements. Comparing our results with those of Shue et al. [37], we find that the web is still small and centralized to a handful of hosting providers. Some IP addresses of major hosting providers host from hundreds of thousands to millions of web sites, which is an increasing trend as these providers often provide web sites hosted on their infrastructure with not only low-cost DoS protection benefits, but also access to their well-provisioned CDN. Regardless of the high degree of web co-location, different from previous observations, our analyses of popular IP block lists show that their collateral damage is relatively small. Since these block lists are carefully curated by reputable organizations, a vast majority of blacklisted IP addresses are associated with only one blocked domain.

2 METHODOLOGY

In this section, we review existing DNS measurement methods and discuss the objectives of our experiment. We also describe how our domain dataset was collected.

2.1 Existing DNS Measurement Techniques

Using *passive measurement*, DNS data is obtained by an entity who is in a position to capture DNS traffic from the network infrastructure under its control (e.g., networks of academic institutes or small organizations) [43]. Several previous studies use passive measurement to observe DNS traffic [5, 8, 20, 37, 43]. Passive measurements, however, may introduce bias in the data collected depending on the time, location, and demographics of users within the monitored network. Moreover, another issue with passive data collection is ethics, as data gathered over a long period of time can reveal online habits of monitored users.

[*] Co-first authors

In contrast, *active measurement* involves sending and receiving DNS queries and responses. Researchers can choose which domains to resolve depending on the goals of their study, thus having more control over the collected data. Although this approach can remedy the privacy issue of passive DNS measurement, it requires an increased amount of resources for running a dedicated measurement infrastructure if there is a large number of domains that need to be resolved [20]. There are prior works that have been conducting large-scale active DNS measurements for different purposes and provide their datasets to the community [20, 33].

However, these datasets have some specific measurement choices that make them unsuitable to be used directly for the purpose of our study. First, all DNS resolutions are issued from a single location (country), while we desire to observe all potential localized IP addresses due to the deployment of CDNs in different regions. Moreover, these datasets aim to exhaustively resolve as many domain names as possible regardless of whether or not they are actively hosting any web content. We further discuss these differences in §4.

2.2 Measurement Objectives

Although it is desirable for us to resolve all web sites to their IP address(es), it is incredibly challenging or even unrealistic to resolve all of them with sufficient regularity (e.g., on a daily basis). As our goal is to study the nature of web co-location and its impact on web users, it is reasonable to focus on active sites that are often visited by the majority of users. To curate such a subset of web sites, we utilized the Alexa and Majestic lists of site rankings. However, only considering the most popular web sites would bias our results. Instead, we tried to include as many sites as possible while keeping our measurements manageable and at the same time, observing a representative subset of web sites on the Internet.

Due to the increasing adoption of load balancing technologies and CDNs, exhaustively resolving *all* possible IP addresses for a given domain can be challenging. To approximate this domain-to-IP mapping, we conducted active DNS measurements from several vantage points obtained from providers of Virtual Private Servers (VPS). We tried to select our measurement locations in a fashion that they are geographically distributed around the globe, thus allowing us to capture as many localized IP addresses of CDN-hosted domains as possible. To that end, we choose nine locations for our measurements, including Brazil, Germany, India, Japan, New Zealand, Singapore, United Kingdom, United States, and South Africa. Our vantage points span the six most populous continents.

2.3 Domain Name Datasets

In the original study, Shue et al. [37] conducted analyses on two datasets of domain names collected from i) the DMOZ Open Directory Project* and ii) the zone files of *.com* and *.net* TLDs. Although it would be ideal to reproduce the study using similar datasets, the DMOZ project was closed in 2017. On the other hand, the number of domains registered under the *.com* and *.net* TLDs has doubled to 156.1M from 75.7M at the time of the original study in 2007.

Pochat et al. [23] recently propose Tranco, which is a list of popular domains combined from data of the most recent 30 days of four top lists that are widely used by the research community: Alexa [2], Majestic [27], Umbrella [40], and Quantcast [31]. However, each top list has its own pitfalls that may negatively impact analysis results if used without careful considerations [23, 35, 36].

Table 1: Daily breakdown of domains and IP addresses observed from each dataset.

	VPS Data	ActiveDNS	Rapid7
Unique domains	8.6M	242M	2B
IPv4-hosted FQDNs	8.2M	117M	1.2B
Unique IPv4s	2.1M	11.5M	710M
IPv6-hosted FQDNs	1.2M	230K	48M
Unique IPv6s	280K	74K	8.8M

To that end, we curated our own domain name dataset from the most recent 30 days of the Alexa and Majestic lists for two reasons. First, these two lists use ranking techniques that are harder and expensive to manipulate [23]. Second, they also have the highest number of common domain names among the four. We do not directly use the Tranco list since it includes domain names from Quantcast and Umbrella. Particularly, Quantcast mostly contains sites that are popular only in the US [23]. Umbrella is highly vulnerable to DNS-based manipulation [23], while it also contains many domains that do not serve web content [35].

From each of our VPS locations, we sent iterative A and AAAA queries for 8.6M fully qualified domain names (FQDNs) on a daily basis. By sending iterative queries, we make sure that local resolvers are not answering these queries from their cache, but the answers come from the actual authoritative name servers. We conducted our measurement for two weeks from July 26th to August 8th, 2019. Our dataset is available at *https://bit.ly/web-colocation-ccr20*.

For comparison, we also repeat our analysis on two public datasets collected during the same period provided by the Active DNS Project [20] and Rapid7 [33] in §4. The Active DNS project queries about 242M domains extracted from zone files of approximately 1.3K TLDs on a daily basis. Rapid7's dataset consists of a much larger number of domains (2B) obtained from zone files, web crawling, and domains returned from PTR records by querying reverse DNS lookup of the whole IPv4 space.

Table 1 summarizes our preliminary observation of unique domains and IP addresses observed in each dataset. Overall, the numbers of domains are much larger than the numbers of IP addresses, indicating that numerous domains are co-hosted under the same IP address(es). We further analyze this co-location degree in §3 and §4.

3 WEB CO-LOCATION ANALYSIS

In this section, we analyze our dataset collected via active DNS measurement to investigate the extent to which web sites are co-located in terms of IP addresses and autonomous systems (AS). We also compare our findings with those found by Shue et al. [37] to examine if previous observations still apply in today's web ecosystem.

*The authors modify the DMOZ dataset to exclude web sites whose domains are in the *.com* and *.net* zone files.

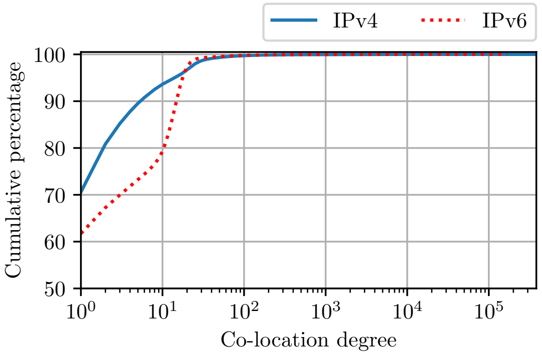


Figure 1: CDF of domains per IP *as a percentage of IPv4/IPv6 addresses* observed in our dataset.

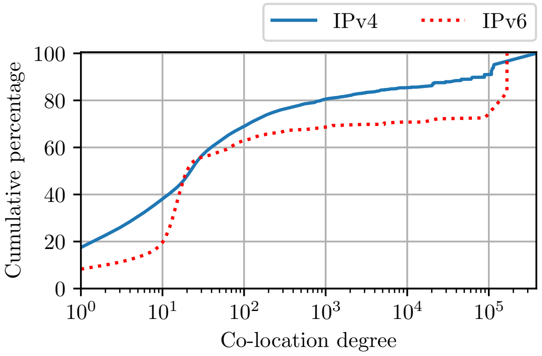


Figure 2: CDF of domains per IP *as a percentage of domains* hosted on IPv4/IPv6 addresses observed in our dataset.

3.1 Web Server Co-location

The co-location degree can be defined in two ways depending on whether we consider an IP address or a domain. When an IP address is considered, the co-location degree is the number of domains hosted on that IP address. Computing the co-location degree of a given domain, however, is more complex as a domain can be hosted on several IP addresses. Therefore, we calculate the co-location degree of a domain by taking the median of co-location degrees across all IP addresses that host that domain.

Figure 1 shows the cumulative distribution function (CDF) of the co-location degree per IP as a percentage of IPv4/IPv6 addresses observed in our dataset. A large portion of both IPv4 (70.5%) and IPv6 (61.7%) are associated with only one domain name. Our findings are similar to those of Shue et al. [37], in which 71% of the IPv4 addresses in their DMOZ dataset host only one domain.

Figure 1 may give an impression that many domains are hosted on their own IP address since there are many IP addresses associated with only one domain. However, according to Figure 2, there are only 17.5% and 8.3% of domains are hosted on their own IPv4 and IPv6 addresses, respectively, without sharing the hosting server with any other domain. This number was higher (24%) in [37] when considering domains obtained from the DMOZ dataset.

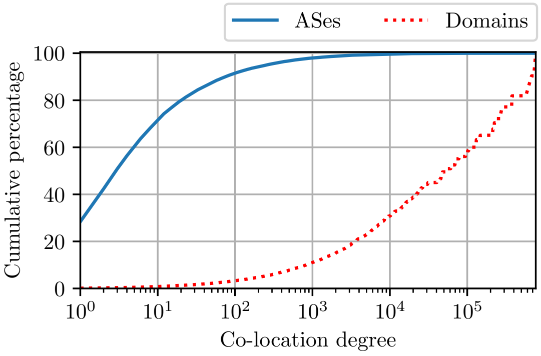


Figure 3: CDF of domains per AS observed in our dataset.

Figure 2 also indicates that about 65% of the web sites in our dataset are co-hosted with 100 or fewer web sites, decreasing from 84% in the study of Shue et al. [37]. On the contrary, we observe that 20% of domains are co-hosted with more than 1K other domains on an IPv4 address, increasing from 6% from the previous study [37]. Our findings show that more domains are co-hosted nowadays.

The other end of the CDF in both Figures 1 and 2 denote a small number of IP addresses having an extremely high degree of co-location, hosting a larger number of domains. The highest co-location degrees are 382K domains for an IPv4 address, and 167K domains for an IPv6 address. Conducting further investigation, we find that the IP address with the highest co-location degree in our dataset belongs to Google, hosting a large number of *blogspot* sites (i.e., sub-sites of *blogger.com*).

3.2 Hosting Provider Co-location

Next, we use CAIDA's *pfx2as* dataset [6] to map IP addresses to their organization (i.e., ASN). Similar to the co-location degree of an IP address (§3.1), the co-location degree in this section is defined as the number of domain names hosted on the same AS.

Figure 3 shows the CDF of domains per AS as a percentage of ASes and domains. 28% of ASes host only one domain while only 0.1% of domains are hosted on an AS themselves. Shue et al. [37] found that there were 60% of domains co-hosted with more than 1K other domains in the same AS. This number has increased to almost 90% of domains as indicated in Figure 3. These findings again show that an even larger number of domains are co-located in terms of their hosting provider, indicating that the web is still small since the first study of Shue et al. conducted more than a decade ago [37].

We further analyze our dataset to investigate which organizations dominate most of the IP addresses and domains. Table 2 shows the top-ten ASes that i) occupy the largest portion of IP addresses, and ii) host the highest number of web sites. As expected, popular CDN providers (e.g., Amazon, Cloudflare, and DigitalOcean) are among the providers from which most IP addresses were observed. However, in terms of the number of domains, Cloudflare and Google are the two largest providers, hosting more than 700K domains each. As Cloudflare provides free web caching services, it is expected to attract many web owners to proxy their web traffic through Cloudflare's CDN. While Google is not among the top-ten ASes that dominate the most IP addresses observed, the company

Table 2: Top hosting providers that have the highest number of IP addresses/domains.

Organization	IPv4s	Organization	Domains
AS16509 Amazon	130K	AS13335 Cloudflare	769K
AS13335 Cloudflare	107K	AS15169 Google	701K
AS14061 DigitalOcean	86K	AS26496 GoDaddy	382K
AS16276 OVH	76K	AS46606 Unified Layer	278K
AS46606 Unified Layer	62K	AS16276 OVH	267K
AS24940 HETZNER-AS	58K	AS16509 Amazon	236K
AS14618 Amazon	57K	AS24940 HETZNER-AS	229K
AS26496 GoDaddy	51K	AS2635 Automattic	145K
AS37963 Alibaba	33K	AS14061 DigitalOcean	130K
AS63949 Linode	31K	AS14618 Amazon	129K

tends to cluster a large number of web sites under a handful of IP addresses, as shown in §3.1. Although a high co-location degree could make co-hosted sites vulnerable to DoS attacks, our finding shows an increasing trend in which more and more web sites are co-hosted and served from well-provisioned CDN of major hosting providers (e.g., Cloudflare, Google). These providers often offer advanced DoS protection benefits at a relatively low cost [7, 16]. A higher co-location degree can also potentially improve the privacy gain of new domain name encryption protocols [19].

Many popular web sites are often served from different IP addresses which may belong to different ASes. We curate our dataset from top lists of popular web sites, and thus are interested in examining whether these web sites are solely hosted on one AS or mirrored on several ASes. More specifically, we examine the top-five populous ASes hosting more than 250K domains, to see if the domains hosted by them are also hosted on other ASes. If a domain is hosted on more than one AS, we call it a "multi-origin" domain.

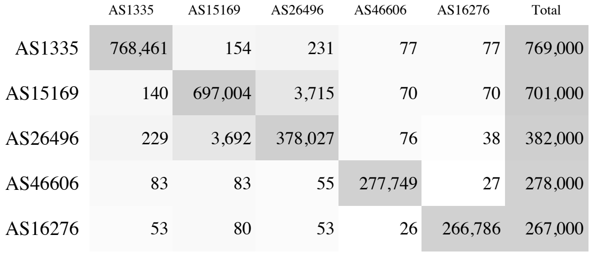
Although Figure 4 shows some multi-origin web sites that are hosted on more than one AS, the number of such web sites is relatively small. More than 99.9% of domains in each AS are only hosted on that AS themselves. This result again confirms that a vast majority of web sites are centralized in a handful of hosting providers. Most web sites are hosted solely on one AS without being mirrored on other ASes. Although most major hosting providers are equipped with enhanced DoS protections, this single-hosting choice may have some impact on the availability of web sites hosted on smaller hosting providers when it comes to targeted DoS attacks.

4 CO-LOCATION DEGREE IN COMPARISON WITH LARGER DNS DATASETS

Next, we repeat our analysis conducted in §3 using two larger DNS datasets to examine the extent to which servers are co-located when considering a much larger number of domain names. More specifically, we analyze the datasets collected by the Active DNS Project [20] and Rapid7 [33] to compare the co-location degree presented in Figures 1, 2, and 3 with these two datasets.

4.1 Dataset Differences

As mentioned in §2, although the two datasets are similar to our dataset in terms of measurement methodology (i.e., active measurement), the location of DNS resolution and the set of resolved domain names are the two reasons making these datasets unsuitable to be used directly for the purpose of our study.



	AS1335	AS15169	AS26496	AS46606	AS16276	Total
AS1335	768,461	154	231	77	77	769,000
AS15169	140	697,004	3,715	70	70	701,000
AS26496	229	3,692	378,027	76	38	382,000
AS46606	83	83	55	277,749	27	278,000
AS16276	53	80	53	26	266,786	267,000

Figure 4: Number of multi-origin domains among top-five autonomous systems. Each cell indicates the number of common domains between two ASes.

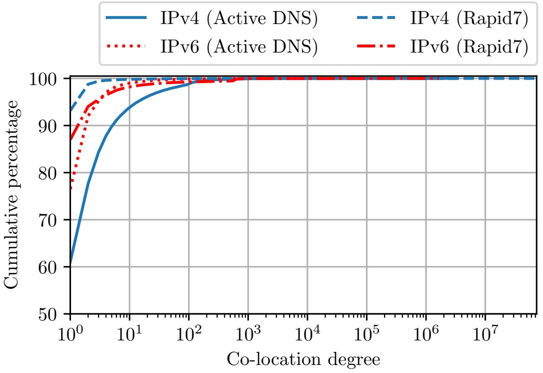


Figure 5: CDF of domains per IP *as a percentage of IPv4/IPv6 addresses* observed in Active DNS and Rapid7 datasets.

With regard to the resolution location, both datasets are collected only from the US. Particularly, the Active DNS dataset is collected at Georgia Tech while the Rapid7 dataset is collected from AWS EC2 nodes in the US. This measurement choice thus could have missed some localized IP addresses of CDN-hosted domains, which we try to obtain by resolving from multiple locations in our experiment.

In terms of the number of resolved domain names, both datasets resolve an order of magnitude larger number of domains than ours as shown in Table 1. Most of these domains, however, are not representative of web sites, while many of them may correspond to spam, phishing [29, 32, 39], malware command and control servers [3], or parking pages registered during the domain drop-catching procedure [4, 22], which most web users would not typically visit. This is the primary reason why we opt to curate our own set of domains from the lists of popular web sites on the Internet.

4.2 Comparison of Co-location Degree

Figure 5 shows the CDF of the co-location degree per IP as a percentage of all IPv4/IPv6 addresses observed in each dataset. Similar to our observation in Figure 1, a large number of both IPv4 and IPv6 addresses are associated with only one domain name. More specifically, 61% of IPv4 addresses in the Active DNS dataset host only one domain, while this number is 93% in the Rapid7 dataset.

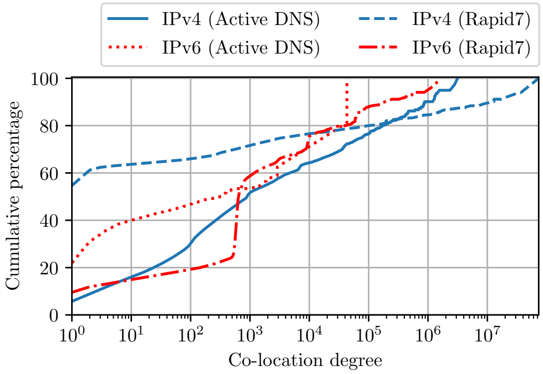


Figure 6: CDF of domains per IP *as a percentage of domains* hosted on IPv4/IPv6 addresses observed in Active DNS and Rapid7 datasets.

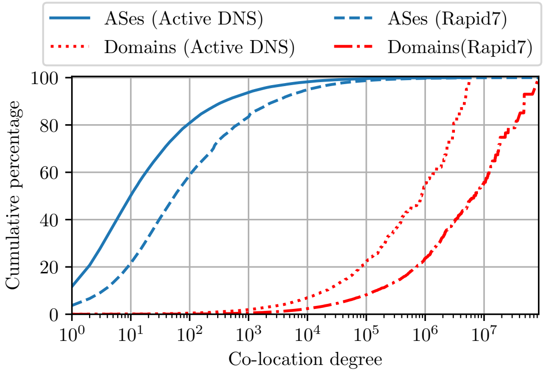


Figure 7: CDF of domains per AS observed in Active DNS and Rapid7 datasets.

This result shows a slight decrease from 69% of IPv4 addresses that host only one domain observed in the study of Shue et al. [37] when considering domains obtained from the *.com* and *.net* zone files.

Similar to our observation in Figure 2, the percentage of domains hosted on an IP themselves is relatively small as shown in Figure 6. The right end of the CDF denotes domains with an extremely high degree of co-location. The highest co-location degrees of Active DNS and Rapid7 are 3.1M and 73.3M per IP address, respectively.

Conducting further investigation, we find that the IP address with the highest co-location degree in Active DNS dataset belongs to Google Cloud and serves more than three million personal and small business domains. For Rapid7, the IP address having the highest co-location degree belongs to AS16276 OVH SAS and hosts more than 73 million mail servers, instead of web content. Understandably, a significant portion of domain names used in Rapid7 dataset consists of PTR records obtained by performing reverse DNS queries over the whole IPv4 address space. Although most reverse DNS lookups do not return a meaningful domain name [19], an IP address hosting an email server is required to have a PTR record, storing the domain name of that email server due to Anti-Spam Recommendations of the Internet Engineering Task Force [24].

Table 3: IP addresses in each block list and common IP addresses between each list and the three DNS datasets.

Block lists	Unique IPs	VPS Data	ActiveDNS	Rapid7
Level1	624,564,857	3,587	34,134	145,540
Level2	35,371	816	1,936	8,175
Level3	37,743	571	1,730	9,747
Level4	9,401,369	21,224	59,948	468,026
Ads	13,422	595	1,895	3,594

Regardless of resolving a much larger number of domain names, Figure 7 shows that only 11.7% of ASes in the Active DNS dataset and 3.7% of ASes in the Rapid7 dataset host one domain, while more than 95% of domains in both datasets are co-hosted on the same AS with at least 10K other domains, showing an extremely high level of AS co-location.

5 BLOCKING COLLATERAL DAMAGE

In this section, we utilize two additional datasets to quantify the collateral damage on co-located domains of IP-based block lists and censorship-motivated block lists.

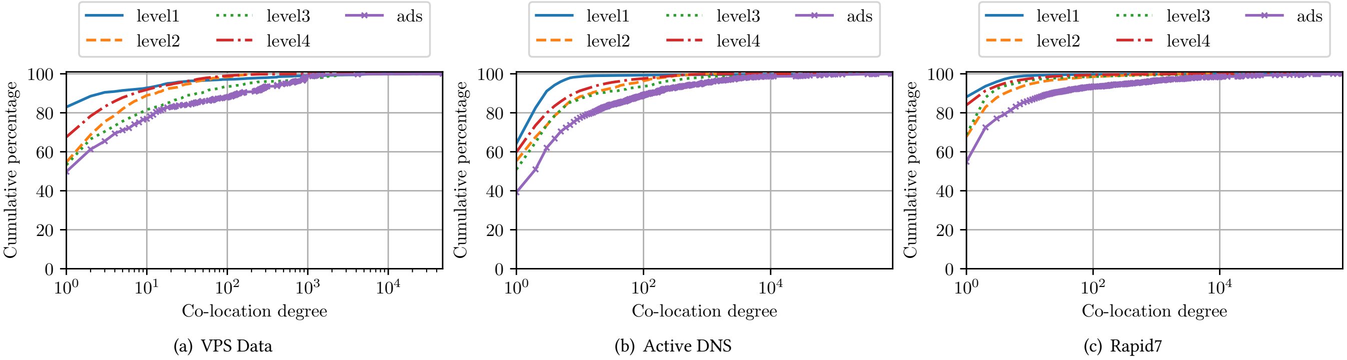
5.1 IP-based blocking collateral damage

In the context of IP-based network filtering, the availability of a web site can be severely impacted by its co-location degree with other sites. For instance, if a powerful DoS attack targets a web site, other co-located sites may also become inaccessible depending on how well-provisioned the hosting infrastructure is. In the initial study by Shue et al. [37], to estimate the collateral damage caused by IP-based blocking, the authors use IP block lists provided by a security company, which unfortunately no longer exists.

For our study, we obtained an additional dataset from FireHOL [15], which is an open-source firewall software that curates its block lists from several highly reputable sources (e.g., *Abuse.ch*, *DShield.org*, and *Spamhaus.org*). Of its block lists, FireHOL aggregates several external well-known lists to create four IP block lists, ranked from 1 to 4, of which *level1* list has minimum false positives, and *level4* may include a large number of false positives.

More specifically, *level1* is curated to include well-known adversarial IP addresses monitored by *Spamhaus.org* and *Team-Cymru.org*. *Level2* includes IP addresses detected to recently conduct brute force attacks. *Level3* contains malicious IP ranges reported by several trustworthy sources in the last 30 days. Finally, *Level4* is made from block lists that track various type of attacks but is susceptible to false positives. In addition, we also utilize FireHOL's lists that contain IP addresses of advertising services and trackers. Table 3 shows the number of IP addresses that each block list has and the number of common IP addresses found in the three DNS datasets used in our study. Of these block lists, *Level1* and *Level4* have the largest numbers of blacklisted IP addresses.

We tested the FireHOL lists, obtained on July 26th, 2019, against the three DNS datasets to estimate the number of domains affected. Figure 8 depicts the CDF of domains affected per blacklisted IP address. Across all three DNS datasets, almost 50% of blacklisted IP addresses host only one domain. As expected, *Level1* has the



(a) VPS Data

(b) Active DNS

(c) Rapid7

Figure 8: CDF of blocked domains per blacklisted IP address as a percentage of common IP addresses between the five block lists and the three DNS datasets.

highest percentage of blacklisted IP addresses that host only one domain, thus causing the least collateral damage. *Level1* is indeed trusted and widely used by the FireHOL community because the list is carefully compiled from well-known sources to minimize false positives. Although *level4* allegedly might include false positives, concerning a high level of collateral damage, its percentage of blacklisted IP addresses hosting only one domain is the second-highest (after *Level1*). Overall, less than 10% of blacklisted IP addresses of these block lists host more than 100 domains. Unlike previous observations, this result shows that state-of-the-art IP block lists are getting better and only cause minimal collateral damage.

5.2 Censorship collateral damage

While domain-name-based blocking is one of the dominant techniques that is often used by censors [1, 9, 11, 17, 30, 38, 42], IP-based blocking can also be very effective for censorship [10, 18, 44]. Currently, domain name information is exposed in either DNS queries or the server name indication (SNI) extension to TLS. This information poses many privacy risks to web users while making it easier for censors to conduct censorship based on the domain name. To remedy these problems, new technologies, including DNS over HTTPS/TLS and ESNI, are introduced to encrypt the domain name information. Under such a circumstance, censors may shift to IP-based blocking if the domain name information cannot be obtained.

To quantify the collateral damage of IP-based censorship, we obtained a list of sensitive sites that are likely to be censored in many countries around the globe. The list is curated by the Citizen Lab [21] and widely used in censorship measurement studies [14, 28]. The list consists of 1,257 web sites, of which we could find 957, 887, and 932 common sites in our dataset, Active DNS, and Rapid7, respectively. We map these domains to their IP address(es) and investigate how many co-located domains would be impacted if a censor conducts IP-based filtering to block these domains.

As indicated in Figure 9, nearly 90% of censored IP addresses host only one sensitive domain, while the highest numbers of affected domains are 11, 6, and 18 for our dataset, Active DNS, and Rapid7, respectively. The result indicates that IP-based censorship will cause little to no collateral damage. To investigate the reason for this finding, we map 2.8K potentially censored IP addresses to their ASN and find 410 unique hosting providers. Of these providers, 280 ASes (68%) host only one domain while 393 ASes (96%) host no more than ten sensitive domains from the Citizen Lab domain list. Therefore, the minimal collateral damage found above is potentially due to the selection of hosting provider used by censored domains. On the other hand, previous actions from the side of providers to hinder domain fronting [13, 34] have shown that the collateral damage [26] caused to hosting providers may have made them unwilling to co-host censored domains with other innocuous domains.

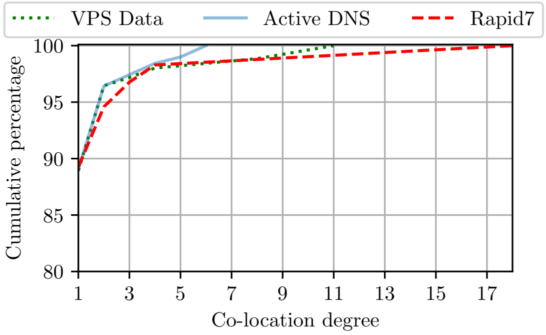


Figure 9: CDF of affected domains per censored IP address as a percentage of all observed IP addresses from censored domains in the Citizen Lab global sensitive list.

6 CONCLUSION

Since its invention, the web has expanded beyond our imagination. More than a decade ago, Shue et al. [37] conducted the first study of web co-location and found that the web was smaller than it seemed. In this paper, we conduct a revisit study of web co-location and could confirm that the web is indeed still small. More specifically, we find that a large number of web sites (often less well-known) are co-hosted on a few IP addresses that belong to major CDN provider. In contrast, a small group of more popular web sites are served from their own well-provisioned servers, occupying a larger pool of IP addresses. While this finding of web co-location is similar to its of Shue et al., our analyses on IP-based blocking show that state-of-the-art IP block lists are getting better, thus causing a very minimal amount of collateral damage.

ACKNOWLEDGEMENTS

We are thankful to Manos Antonakakis, Panagiotis Kintis, and Logan O'Hara from the Active DNS Project for providing us their DNS dataset, and to Rapid7 for making their datasets available to the research community. We also thank all the anonymous reviewers for their thorough feedback on earlier drafts of this paper.

This research was supported in part by the National Science Foundation under awards CNS-1740895 and CNS-1719386. The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the sponsors.

REFERENCES

[1] 2014. Towards a Comprehensive Picture of the Great Firewall's DNS Censorship. In *4th Workshop on Free and Open Communications on the Internet*. USENIX.

[2] Alexa Internet, Inc. Accessed 2019. Alexa Top Global Sites. https://alexa.com/

[3] Eihal Alowaisheq, Peng Wang, Sumayah Alrwais, Xiaojing Liao, XiaoFeng Wang, Tasneem Alowaisheq, Xianghang Mi, Siyuan Tang, and Baojun Liu. 2019. Cracking the Wall of Confinement: Understanding and Analyzing Malicious Domain Take-downs. In *Network and Distributed System Security*. Internet Society.

[4] Timothy Barron, Najmeh Miramirkhani, and Nick Nikiforakis. 2019. Now You See It, Now You Don't: A Large-scale Analysis of Early Domain Deletions. In *22nd International Symposium on Research in Attacks, Intrusions and Defenses*.

[5] Leyla Bilge, Engin Kirda, Christopher Kruegel, and Marco Balduzzi. 2011. EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis. In *Network and Distributed System Security Symposium*.

[6] Center for Applied Internet Data Analysis. Accessed 2019. Routeviews Prefix to AS mappings Dataset for IPv4 and IPv6 . Web page. http://www.caida.org/data/routing/routeviews-prefix2as.xml

[7] Cloudflare. Accessed 2019. How does Cloudflare work? Web page. https://support.cloudflare.com/hc/en-us/articles/205177068-How-does-Cloudflare-work-

[8] Matteo Dell'Amico, Leyla Bilge, Ashwin Kayyoor, Petros Efstathopoulos, and Pierre-Antoine Vervier. 2017. Lean On Me: Mining Internet Service Dependencies From Large-Scale DNS Data. In *Proceedings of the 33rd Annual Computer Security Applications Conference (ACSAC 2017)*. ACM, New York, NY, USA, 449–460.

[9] Hai-Xin Duan, Nicholas Weaver, Zengzhi Zhao, Meng Hu, Jinjin Liang, Jian Jiang, Kang Li, and Vern Paxson. 2012. Hold-On: Protecting Against On-Path DNS Poisoning. In *the Conference on Securing and Trusting Internet Names (SATIN)*.

[10] Arun Dunna, Ciarán O'Brien, and Phillipa Gill. 2018. Analyzing China's Blocking of Unpublished Tor Bridges. In *8th USENIX Workshop on Free and Open Communications on the Internet*. USENIX, Baltimore, MD.

[11] Oliver Farnan, Alexander Darer, and Joss Wright. 2016. Poisoning the Well: Exploring the Great Firewall's Poisoned DNS Responses. In *Workshop on Privacy in the Electronic Society*. ACM, New York, 95–98.

[12] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, and T. Berners-Lee. 1997. *Hypertext Transfer Protocol – HTTP/1.1*. RFC 2068. IETF. https://tools.ietf.org/html/rfc2068

[13] David Fifield, Chang Lan, Rod Hynes, Percy Wegmann, and Vern Paxson. 2015. Blocking-resistant communication through domain fronting. *Proceedings on Privacy Enhancing Technologies* 2015, 2 (2015), 46–64.

[14] Arturo Filastò and Jacob Appelbaum. 2012. OONI: Open Observatory of Network Interference. In *Free and Open Communications on the Internet*. USENIX. https://www.usenix.org/system/files/conference/foci12/foci12-final12.pdf

[15] FireHOL. Accessed 2019. All Cybercrime IP Feeds. Web page. http://iplists.firehol.org

[16] Google. Accessed 2019. Best Practices for DDoS Protection and Mitigation on Google Cloud Platform. Web page. http://cloud.google.com/files/GCPDDoSprotection-04122016.pdf

[17] Nguyen Phong Hoang, Sadie Doreen, and Michalis Polychronakis. 2019. Measuring I2P Censorship at a Global Scale. In *9th USENIX Workshop on Free and Open Communications on the Internet (FOCI 19)*. USENIX Association, Santa Clara, CA.

[18] Nguyen Phong Hoang, Panagiotis Kintis, Manos Antonakakis, and Michalis Polychronakis. 2018. An Empirical Study of the I2P Anonymity Network and Its Censorship Resistance. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, New York, NY, USA, 379–392.

[19] Nguyen Phong Hoang, Arian Akhavan Niaki, Nikita Borisov, Phillipa Gill, and Michalis Polychronakis. 2020. Assessing the Privacy Benefits of Domain Name Encryption. In *Proceedings of the 15th ACM ASIA Conference on Computer and Communications Security (ASIACCS '20)*. ACM, New York, NY, USA. https://doi.org/10.1145/3320269.3384728

[20] Athanasios Kountouras, Panagiotis Kintis, Chaz Lever, Yizheng Chen, Yacin Nadji, David Dagon, Manos Antonakakis, and Rodney Joffe. 2016. Enabling Network Security Through Active DNS Datasets. In *Research in Attacks, Intrusions, and Defenses*, Fabian Monrose, Marc Dacier, Gregory Blanc, and Joaquin Garcia-Alfaro (Eds.). Springer International Publishing, Cham, 188–208.

[21] Citizen Lab and Others. 2014. URL testing lists intended for discovering website censorship. https://github.com/citizenlab/test-lists https://github.com/citizenlab/test-lists.

[22] Tobias Lauinger, Abdelberi Chaabane, Ahmet Salih Buyukkayhan, Kaan Onarlioglu, and William Robertson. 2017. Game of Registrars: An Empirical Analysis of Post-Expiration Domain Name Takeovers. In *26th USENIX Security Symposium (USENIX Security 17)*. USENIX Association, Vancouver, BC, 865–880. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/lauinger

[23] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019)*. https://doi.org/10.14722/ndss.2019.23386

[24] G. Lindberg. 1999. *Anti-Spam Recommendations for SMTP MTAs*. RFC 2505. IETF. https://tools.ietf.org/html/rfc2505

[25] Netcraft Ltd. Accessed 2019. How many active sites are there? Web page. https://www.netcraft.com/active-sites/

[26] Neil MacFarquhar. 2018. Russia Tried to Shut Down Telegram. Websites Were Collateral Damage. https://www.nytimes.com/2018/04/18/world/europe/russia-telegram-shutdown.html.

[27] Majestic. Accessed 2019. The Majestic Million. Web page. https://majestic.com/reports/majestic-million

[28] Arian Akhavan Niaki, Shinyoung Cho, Zachary Weinberg, Nguyen Phong Hoang, Abbas Razaghpanah, Nicolas Christin, and Phillipa Gill. 2020. ICLab: A Global, Longitudinal Internet Censorship Measurement Platform. In *IEEE Symposium on Security and Privacy*.

[29] Elkana Pariwono, Daiki Chiba, Mitsuaki Akiyama, and Tatsuya Mori. 2018. Don'T Throw Me Away: Threats Caused by the Abandoned Internet Resources Used by Android Apps. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security (ASIACCS '18)*. ACM, New York, NY, USA, 147–158. https://doi.org/10.1145/3196494.3196554

[30] Paul Pearce, Ben Jones, Frank Li, Roya Ensafi, Nick Feamster, Nick Weaver, and Vern Paxson. 2017. Global Measurement of DNS Manipulation. In *26th USENIX Security Symposium*.

[31] quantcast. Accessed 2019. Quantcast Top Websites. Web page. https://www.quantcast.com/top-sites/

[32] Florian Quinkert, Tobias Lauinger, William Robertson, Engin Kirda, and Thorsten Holz. 2019. It's Not What It Looks Like: Measuring Attacks and Defensive Registrations of Homograph Domains. In *2019 IEEE Conference on Communications and Network Security (CNS)*.

[33] Rapid7. 2019. Rapid7: Open Data. https://opendata.rapid7.com/.

[34] Fahmida Y. Rashid. 2018. Amazon joins Google in shutting down doamin fronting. https://duo.com/decipher/amazon-joins-google-in-shutting-down-domain-fronting.

[35] Walter Rweyemamu, Christo Lauinger, Tobiasand Wilson, William Robertson, and Engin Kirda. 2019. Clustering and the Weekend Effect: Recommendations for the Use of Top Domain Lists in Security Research. In *Passive and Active Measurement*, David Choffnes and Marinho Barcellos (Eds.). Springer International Publishing, 161–177.

[36] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, New York, NY, USA, 478–493. https://doi.org/10.1145/3278532.3278574

[37] Craig A. Shue, Andrew J. Kalafut, and Minaxi Gupta. 2007. The Web is Smaller Than It Seems. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC '07)*. ACM, New York, NY, USA, 123–128. https://doi.org/10.1145/1298306.1298324

[38] Sparks, Neo, Tank, Smith, and Dozer. 2012. The Collateral Damage of Internet Censorship by DNS Injection. *SIGCOMM Computer Communication Review* 42, 3 (2012), 21–27. http://conferences.sigcomm.org/sigcomm/2012/paper/ccr-paper266.pdf

[39] Ke Tian, Steve T. K. Jan, Hang Hu, Danfeng Yao, and Gang Wang. 2018. Needle in a Haystack: Tracking Down Elite Phishing Domains in the Wild. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, New York, NY, USA, 429–442. https://doi.org/10.1145/3278532.3278569

[40] Cisco Umbrella. Accessed 2019. Umbrella Popularity List. Web page. https://s3-us-west-1.amazonaws.com/umbrella-static/index.html

[41] Verisign. 2019. *The Domain Name Industry Brief*. Technical Report. Verisign. https://www.verisign.com/assets/domain-name-report-Q22019.pdf

[42] M. Wander, C. Boelmann, L. Schwittmann, and T. Weis. 2014. Measurement of Globally Visible DNS Injection. *IEEE Access* 2 (2014), 526–536.

[43] Florian Weimer. 2005. Passive DNS replication. In *FIRST conference on computer security incident*. 98.

[44] Philipp Winter and Stefan Lindskog. 2012. How the Great Firewall of China is Blocking Tor. In *Free and Open Communications on the Internet*. USENIX. https://www.usenix.org/system/files/conference/foci12/foci12-final2.pdf