Assessing the Privacy Benefits of Domain Name Encryption

Nguyen Phong Hoang
Stony Brook University
nghoang@cs.stonybrook.edu

Arian Akhavan Niaki
University of Massachusetts, Amherst
arian@cs.umass.edu

Nikita Borisov
University of Illinois at Urbana-Champaign
nikita@illinois.edu

Phillipa Gill
University of Massachusetts, Amherst
phillipa@cs.umass.edu

Michalis Polychronakis
Stony Brook University
mikepo@cs.stonybrook.edu

ABSTRACT

As Internet users have become more savvy about the potential for their Internet communication to be observed, the use of network traffic encryption technologies (e.g., HTTPS/TLS) is on the rise. However, even when encryption is enabled, users leak information about the domains they visit via DNS queries and via the Server Name Indication (SNI) extension of TLS. Two recent proposals to ameliorate this issue are DNS over HTTPS/TLS (DoH/DoT) and Encrypted SNI (ESNI). In this paper we aim to assess the privacy benefits of these proposals by considering the relationship between hostnames and IP addresses, the latter of which are still exposed. We perform DNS queries from nine vantage points around the globe to characterize this relationship. We quantify the privacy gain offered by ESNI for different hosting and CDN providers using two different metrics, the *k*-anonymity degree due to co-hosting and the dynamics of IP address changes. We find that 20% of the domains studied will not gain any privacy benefit since they have a one-to-one mapping between their hostname and IP address. On the other hand, 30% will gain a significant privacy benefit with a *k* value greater than 100, since these domains are co-hosted with more than 100 other domains. Domains whose visitors' privacy will meaningfully improve are far less popular, while for popular domains the benefit is not significant. Analyzing the dynamics of IP addresses of long-lived domains, we find that only 7.7% of them change their hosting IP addresses on a daily basis. We conclude by discussing potential approaches for website owners and hosting/CDN providers for maximizing the privacy benefits of ESNI.

CCS CONCEPTS

• **Networks** → **Network privacy and anonymity**; **Network measurement**;

KEYWORDS

Domain name privacy; DNS over HTTPS (DoH); DNS over TLS (DoT); Encrypted SNI (ESNI); active DNS measurement.

ASIA CCS '20, October 5–9, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-6750-9/20/10…$15.00
https://doi.org/10.1145/3320269.3384728

ACM Reference Format:
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 (ASIA CCS '20), October 5–9, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3320269.3384728

1 INTRODUCTION

As users become more aware of the importance of protecting their online communication, the adoption of TLS is increasing [60]. It is indicative that almost 200M fully-qualified domain names (FQDNs) support TLS [38], while Let's Encrypt [3] has issued a billion certificates as of February 27, 2020 [4]. Although TLS significantly improves the confidentiality of Internet traffic, on its own it cannot fully protect user privacy, especially when it comes to monitoring the websites a user visits.

Currently, visited domain names are exposed in both i) DNS requests, which remain unencrypted, and ii) the Server Name Indication (SNI) extension [56] during the TLS handshake. As a result, on-path observers can fully monitor the domain names visited by web users through simple eavesdropping of either DNS requests or TLS handshake traffic. Several recent proposals aim to improve the security and privacy of these two protocols. Specifically, DNS over HTTPS (DoH) [53] and DNS over TLS (DoT) [55] aim to preserve the integrity and confidentiality of DNS resolutions against threats "below the recursive," such as DNS poisoning [32], while Encrypted Server Name Indication (ESNI) [89] aims to prevent "nosy" ISPs and other on-path entities from observing the actual visited domain of a given TLS connection.

In this paper, we quantify the potential improvement to user privacy that a full deployment of DoH/DoT and ESNI would achieve in practice, given that destination IP addresses still remain visible to on-path observers. Although it is straightforward to reveal a user's visited site if the destination IP address hosts only that particular domain, when a given destination IP address serves many domains, an adversary will have to "guess" which one is being visited.

We use two properties to quantify the potential privacy benefit of ESNI, assuming the provider of the DoH/DoT server used is fully trusted (as it can still observe all visited domains): the *k*-anonymity property and the dynamics of hosting IP addresses. Assuming that *k* different websites are co-hosted on a given IP address (all using HTTPS with ESNI supported), the privacy of a visitor to one of those sites increases as the number of *k-1* other co-hosted sites increases. In addition, the more dynamic the hosting IP address is

for a given site, the higher the privacy benefit of its visitors is, as the mapping between domain and hosting IP address becomes less stable, and thus less predictable.

To quantify these two properties, we conducted active DNS measurements to obtain the IP addresses of an average of 7.5M FQDNs per day drawn from lists of popular websites [7, 70] (§4). To account for sites served from content delivery networks (CDNs) which may direct users differently based on their location, we performed name resolutions from nine locations around the world: Brazil, Germany, India, Japan, New Zealand, Singapore, United Kingdom, United States, and South Africa. Our measurements were conducted in two months to investigate how much a network observer can learn about the domains visited by a user based solely on the IP address information captured from encrypted traffic.

We find that 20% of the domains studied will not benefit at all from ESNI, due to their stable one-to-one mappings between domain name and hosting IP address. For the rest of the domains, only 30% will gain a significant privacy benefit with a k value greater than 100, which means an adversary can correctly guess these domains with a probability lower than 1%. The rest 50% of the domains can still gain some privacy benefits, but at a lower level (i.e., $2 \leq k \leq 100$). While sophisticated website fingerprinting attacks based on several characteristics of network packets (e.g., timing and size [45, 68, 74, 78, 79, 101]) can be used to predict the visited domains, our study aims to provide a lower bound of what an attacker can achieve.

Moreover, we observe that sites hosted by the top-ten hosting providers with the highest privacy value ($k > 500$) are far less popular (§5.2). These are often less well-known sites hosted on small hosting providers that tend to co-locate many websites on a single IP or server. In contrast, the vast majority of more popular sites would gain a much lower level of privacy. These sites are often hosted by major providers, including Cloudflare ($k = 16$), Amazon ($3 \leq k \leq 5$), Google ($k = 5$), GoDaddy ($k = 4$), and Akamai ($k = 3$).

In addition, we find that frequently changing IP addresses (at least once a day) are limited to only 7.7% of the domains that we were able to resolve each day of our study. As expected, dominant providers in terms of more dynamic IP addresses include major CDN providers, such as Amazon, Akamai, and Cloudflare (§5.4).

Finally, we validate and compare our main findings by repeating part of our analysis using two different public DNS datasets (§6), and provide recommendations for both website owners and hosting/CDN providers on how to maximize the privacy benefit offered by the combination of DoH/DoT and ESNI (§7). In particular, website owners may want to seek hosting services from—the unfortunately quite few—providers that maximize the ratio between co-hosted domains per IP address, and minimize the duration of domain-to-IP mappings. Hosting providers, on the other hand, can hopefully aid in maximizing the privacy benefits of ESNI by increasing the unpredictability of domain-to-IP mappings.

2 BACKGROUND

In this section, we provide an overview of the shortcomings of DNS and TLS when it comes to user privacy, along with the suggested improvements of DNS over HTTPS/TLS (DoH/DoT) and Encrypted Server Name Indication (ESNI).

2.1 DNS and DoH/DoT

The DNS protocol exposes all requests and responses in plaintext, allowing anyone with the privilege to monitor or modify a user's network traffic to eavesdrop or tamper with it. For example, a man-on-the-side attacker can send spoofed DNS responses to redirect a victim to malicious websites [32], while state-level organizations can manipulate DNS responses to disrupt connections for censorship purposes. The DNS Security Extensions (DNSSEC) [36] were introduced in 1997 to cope with these and other security issues by assuring the integrity and authenticity of DNS responses (but not their confidentiality). However, DNSSEC is still not widely deployed due to deployment difficulties and compatibility issues [23, 29, 44].

As an attempt to enhance the security and privacy of the DNS protocol, two emerging DNS standards were recently proposed: DoH [53] and DoT [55]. These technologies aim to not only ensure the integrity and authenticity of DNS traffic, but also its confidentiality to some extent. Using DoH/DoT, all DNS queries and responses are transmitted over TLS, ensuring their integrity against last-mile adversaries who would otherwise be in a position to launch man-in-the-middle (MiTM) and man-on-the-side (MoTS) attacks. In this work, we specifically characterize the protection of user privacy from nosy ISPs and other last-mile entities provided by DoH/DoT.

Although the benefits of DoH/DoT against last-mile adversaries are clear, this comes with the cost of *fully trusting* a third-party operator of the DoH/DoT resolver on which users have outsourced all their DNS resolutions [51]. Several companies already offer public DoH/DoT resolvers, including Google [42, 43] and Cloudflare [108]. In fact, we later show in §5.2 that these two companies are also the most dominant *hosting* providers of domains in the top lists of popular sites. Popular browsers have also started introducing support for DoH/DoT, e.g., Mozilla Firefox since version 62 [72] (which is now enabled by default).

2.2 SSL/TLS and ESNI

During the TLS handshake [31], the two communicating parties exchange messages to acknowledge and verify the other side using digital certificates, and agree on various parameters that will be used to create an encrypted channel. In a client-server model, the client trusts a digital certificate presented by the server as long as it has been signed by a trusted certificate authority.

Ideally, private or sensitive information should be transmitted only after the TLS handshake has completed. This goal can be easily achieved when a server hosts only a single domain (known as IP-based virtual hosting). Name-based virtual hosting, however, which is an increasingly used approach for enabling multiple domains to be hosted on a *single* server, necessitates a mechanism for the server to know which domain name a user intends to visit *before* the TLS handshake completes, in order to present the right certificate. The Server Name Indication (SNI) extension was introduced in 2003 as a solution to this problem. The SNI extension contains a field with the domain name the client wants to visit, so that the server can then present the appropriate certificate. Unfortunately, since this step is conducted prior to the completion of the TLS handshake, the domain name specified in SNI is exposed in plaintext. Consequently, all the privacy risks associated with the traditional design of DNS discussed above also apply to the SNI extension. For

instance, Internet authorities in several countries have been relying on the SNI field for censorship purposes [21, 49].

ESNI has recently been proposed as part of TLS version 1.3 [89] to resolve the issue of SNI revealing the domain visited by a user. Using ESNI, clients encrypt the SNI field towards a given server by first obtaining a server-specific public key through a well-known ESNI DNS record. Obviously, due to this reliance on DNS, any privacy benefits of ESNI can only be realized when used in conjunction with DoH/DoT—otherwise any last-mile observer would still be able to observe a user's plaintext DNS queries and infer the visited TLS server. In September 2018, Cloudflare was among the first providers to announce support for ESNI across its network [83].

3 THREAT MODEL

We assume an idealistic future scenario in which both DoH/DoT and ESNI are *fully* deployed on the Internet. Under this assumption, an on-path observer will only be able to rely on the remaining visible information, i.e., destination IP addresses, to infer the sites being visited by the monitored users. The extent to which this inference can be easily made depends on i) whether other domains are hosted on the same IP address, and ii) the stability of the mapping between a given domain and its IP address(es).

The probability with which an adversary can successfully infer the visited domain can be modeled using the k-anonymity property, with k corresponding to the number of domains co-hosted on the same IP address. The probability of a successful guess is inversely proportional to the value of k, i.e., the larger the k, the more difficult it is for the adversary to make a correct guess, thus providing increased user privacy.

The above threat model is oblivious to distinguishable characteristics among a group of co-hosted websites, such as popularity ranking, site sensitivity, and network traffic patterns. We should thus stress that the situation in practice will be *much more favorable* for the adversary. Even for a server with a high k, it is likely that not all k sites will be equally popular or sensitive. Although the popularity and sensitivity can vary from site to site, depending on who, when, and from where is visiting the site [48], an adversary can still consider the popularity and sensitivity of the particular k sites hosted on a given IP address to make a more educated guess about the actual visited site.

Utilizing the ranking information of all domains studied, we model such an adversarial scenario in §5.3 and show that our threat model based on k-anonymity is still valid. In addition, page-specific properties such as the number of connections towards different third-party servers and the number of transferred bytes per connection can be used to derive robust web page *fingerprints* [16, 17, 27, 35, 47, 65, 81, 99], which can improve the accuracy of attribution even further. Although identifying a visited website among all possible websites on the Internet by relying solely on fingerprinting is quite challenging, applying the same fingerprinting approach for attributing a given connection (and subsequent associated connections) to one among a set of k *well-known* websites is a vastly easier problem.

Consequently, an on-path observer could improve the probability of correctly inferring the actual visited website by considering the popularity and sensitivity of the co-hosted domains on the visited IP address, perhaps combined with a form of traffic fingerprinting. Although such a more powerful attack is outside the scope of this work, as we show in the rest of the paper, our results already provide a worrisome insight on how effective an even much less sophisticated attribution strategy would be, given the current state of domain co-hosting.

4 METHODOLOGY

In this section we review existing DNS measurement techniques and highlight the data collection goals of our study. We then describe how we select domains and vantage points to achieve these goals.

4.1 Existing DNS Measurements

Previous studies use passive measurement to observe DNS traffic on their networks [30, 94, 100]. However, passive data collection can suffer from bias depending on the time, location, and demographics of users within the observed network. Passive data collection can also raise ethical concerns, as data collected over a long period of time can gradually reveal online habits of monitored users.

There are also prior works (by both academia and industry) that conducted large-scale active DNS measurements for several purposes and made their datasets available to the community [61, 87]. However, these datasets have two common issues that make them unsuitable to be used directly in our study. First, all DNS queries are resolved from a single location (country), while we aim to observe localized IPs delivered by CDNs to users in different regions. Second, although these datasets have been used in many other studies, none of the prior measurements are designed to filter out poisoned DNS responses (e.g., as a result of censorship leakage), which can significantly affect the accuracy of the results and negatively impact data analysis if not excluded. We discuss steps taken to sanitize these datasets in Appendix B.

4.2 Our Measurement Goals

Ideally, we would like to derive the mapping between all live domain names and their IP addresses. Unfortunately, this is extremely challenging to achieve in practice because there are more than 351.8 million second-level domain names registered across all top-level domains (TLDs) at the time we compose this paper [98], making it unrealistic to actively resolve all of them with adequate frequency. Furthermore, not all domains host web content, while many of them correspond to spam, phishing [80, 85], malware command and control [8], or parking pages registered during the domain dropcatching process [63], which most users do not normally visit.

As we aim to study the privacy benefits of ESNI, we thus choose to focus on active sites that are legitimately visited by the majority of web users. To derive such a manageable subset of sites, we relied on lists of website rankings, but did not consider only the most popular ones, as this would bias our results. Instead, we expanded our selection to include as many sites as possible, so that we can keep our measurements manageable, but at the same time observe a representative subset of *legitimately visited* domains on the Internet.

4.3 Domains Tested

There are four top lists that are widely used by the research community: Alexa [7], Majestic [70], Umbrella [97], and Quantcast [84].

However, it is challenging to determine which top list should be chosen, as recent works have shown that each top list has its own issues that may significantly affect analysis results if used without some careful considerations [64, 90, 92]. For instance, Alexa is highly fluctuating, with more than 50% of domain names in the list changing every day, while Majestic is more stable but cannot quickly capture sites that suddenly become popular for only a short period of time. Pochat et al. [64] suggest that researchers should combine these four lists to generate a reliable ranking.

For this study, we generated our own list by aggregating domains ranked by Alexa and Majestic from the most recent 30 days for several reasons. First, these two lists use ranking techniques that are more difficult and costly to manipulate [64]. Second, they have the highest number of domains in common among the four. We exclude domains from Quantcast because it would make our observations biased towards popular sites only in the US [64]. Lastly, we do not use domains from Umbrella because the list is vulnerable to DNS-based manipulation and also contains many domains that do not host web content [64, 90]. To this end, we studied a total of 13.6M domains with its breakdown shown in Appendix A.

Data scope. Although this subset of domains corresponds to about 4% of all domains in the TLD zone files, we argue that it is still adequate for the goal of our study, i.e., determining whether the current state of website co-location will allow ESNI to provide a meaningful privacy benefit. Considering only this subset of domains may lead to an under-approximation of the actual k-anonymity offered by a given IP or set of IPs, as some co-hosted domains may not be considered. This means that our results can be viewed as a lower bound of the actual k-anonymity degree for a given visited IP address, which is still a desirable outcome.

As discussed in §3, the popularity of a website, along with other qualitative characteristics, can be used by an adversary to improve attribution. Indeed, given that the long tail of domains that are left out from our dataset mostly correspond to vastly less popular and even unwanted or dormant domains [96], any increase in k they may contribute would in practice be rather insignificant, as (from an attribution perspective) it is unlikely they will be the ones that most web users would actually visit.

4.4 Measurement Location and Duration

Due to load balancing and content delivery networks, deriving *all* possible IP addresses for a given popular domain is very challenging. To approximate this domain-to-IP mapping, we performed our own active DNS measurements from several vantage points acquired from providers of Virtual Private Servers (VPS). When choosing measurement locations, we tried to distribute our vantage points so that their geographical distances are maximized from each other. This design decision allows us to capture as many localized IP addresses of CDN-hosted sites as possible. To that end, we run our measurements from nine countries, including Brazil, Germany, India, Japan, New Zealand, Singapore, United Kingdom, United States, and South Africa. Our vantage points span the six most populous continents. From all measurement locations mentioned above, we send DNS queries for approximately 7.5M domains on a daily basis. When issuing DNS queries, we enabled the iterative flag in the queries, bypassing local recursive resolvers to make sure that DNS responses are returned by actual authoritative name servers. The results presented in this work are based on data collected for a period of two months, from February 24th to April 25th, 2019.

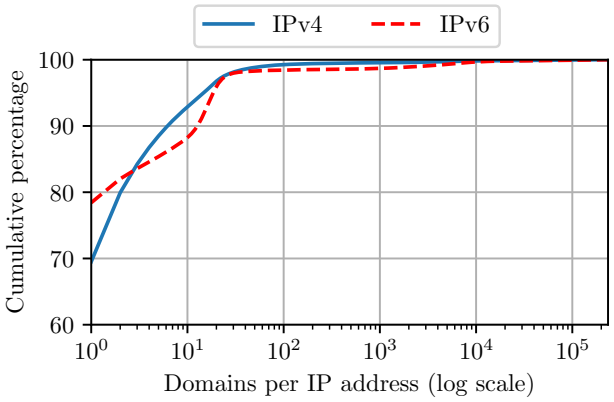

Figure 1: Cumulative distribution function (CDF) of the number of domains hosted per IP address, as a percentage of all observed IP addresses. About 70% of all observed IPv4 addresses host only a single domain.

5 DATA ANALYSIS

In this section we use two metrics, k-anonymity and the dynamics of hosting IP addresses, to quantify the privacy benefits offered by different hosting and CDN providers. To verify the validity of our k-anonymity model, we also apply Zipf's law on the popularity ranking of domains to account for a more realistic (i.e., more powerful) adversary.

5.1 Single-hosted vs. Multi-hosted Domains

Over a period of two months, from February 24th to April 25th, 2019, we observed an average of 2.2M and 500K unique IPv4 and IPv6 addresses, respectively, from our daily measurements. Of these IP addresses, 70% of IPv4 and 79% of IPv6 addresses host only a single domain, as shown in Figure 1. This means that visitors of the websites hosted on those addresses will not gain any meaningful privacy benefit with ESNI, due to the one-to-one mapping between the domain name and the IP address on which it is hosted. About 95% of both IPv4 and IPv6 addresses host less than 15 domains.

When calculating the percentage of IPv6-supported sites, we find that less than 15% support IPv6. Regardless of the increasing trend [28], the future adoption of IPv6 is still unclear [25]. Since the majority of web traffic is still being carried through IPv4, in the rest of the paper we focus only on IPv4 addresses.

Based on our measurements, we identify three main ways in which a domain may be hosted, in terms of the IP addresses used and the potential privacy benefit due to ESNI, as illustrated in Figure 2. In the simplest case, a *single-hosted* domain may be *exclusively hosted* on one or more IP addresses that do not serve any other domain, to which we refer as *privacy-detrimental* IP addresses (Fig. 2, left). As there is no sharing of the IP address(es) with other domains, an adversary can trivially learn which site is visited based solely on the destination IP address. On the other hand, a *multi-hosted*

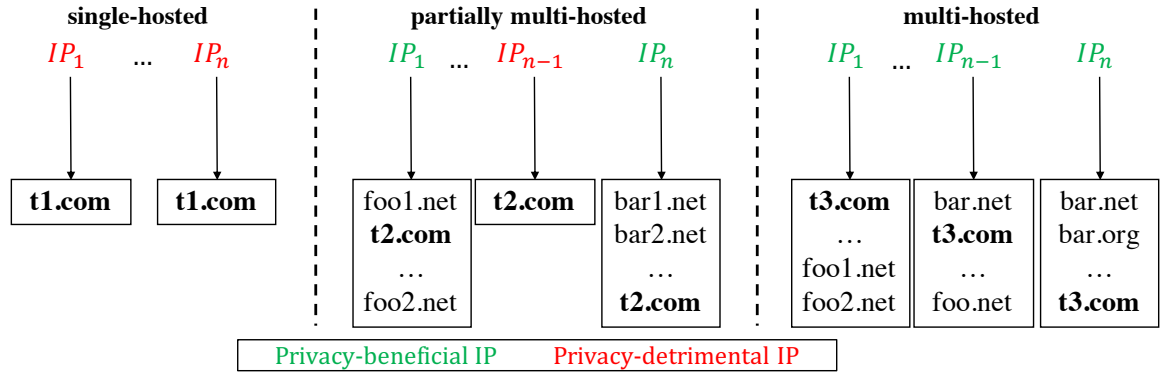

Figure 2: Different types of domain hosting according to whether they can benefit from ESNI. Single-hosted domains are exclusively hosted on one or more IP addresses, and thus cannot benefit from ESNI. In contrast, multi-hosted domains are always co-hosted with more domains on a given IP address, and thus can benefit by ESNI.

domain (Fig. 2, right) may be *co-hosted* on one or more IP addresses that always serve at least one or more other domains, to which we refer as *privacy-beneficial* IP addresses. Since the destination IP address always hosts multiple domains, an adversary can only make a (possibly educated) guess about the actual domain a given user visits, and thus multi-hosted domains always benefit to some extent from ESNI—the more co-hosted domains on a given IP address, the higher the privacy gain offered by ESNI.

Finally, there is a chance that a domain is hosted on a mix of privacy-detrimental and privacy-beneficial IP addresses, which we call *partially multi-hosted* domains (Fig. 2, middle). In that case, only visitors to the subset of IP addresses that co-host other domains will benefit from ESNI. Based on our measurements, partially multi-hosted domains correspond to only a 0.3% fraction (20K) of all domains (daily average). Single-hosted domains comprise 18.7% (1.4M) and multi-hosted domains comprise 81% (6M) of all domains.

The privacy degree of a partially multi-hosted domain depends on the probability that a visitor gets routed to a privacy-beneficial IP of that domain. In other words, a partially multi-hosted domain will mostly behave as a multi-hosted domain if the majority of its IP addresses are privacy-beneficial. In fact, we find that this is the case for more than 92.5% of the partially multi-hosted domains studied. Based on this fact, and given its extremely small number compared to the other two types, in the rest of our paper we merge partially multi-hosted domains with the actual multi-hosted domains, to simplify the presentation of our results.

Going back to Figure 1, based on the above breakdown, we observe that 70% of all IP addresses that host a single domain correspond to 18.7% of all domains, i.e., the single-hosted ones. On the other hand, the 81% of multi-hosted domains are co-hosted on just 30% of the IP addresses observed.

Next, we analyze the popularity distribution of single-hosted and multi-hosted domains to identify any difference in the user population of these two types of domains. Note that we only base our analysis on the ranking information provided by the top lists to comparatively estimate the scale of the user base, and not for absolute ranking purposes. More specifically, we only use the top 100K domains for the analysis in Figure 3, since rankings lower

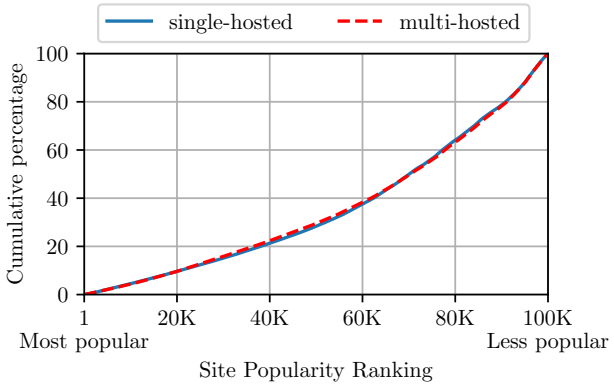

Figure 3: CDF of the popularity ranking for single-hosted and multi-hosted domains.

than 100K are not statistically significant, as confirmed by both top list providers and recent studies [6, 90]. Figure 3 shows that single-hosted and multi-hosted domains exhibit a nearly identical distribution of popularity rankings.

5.2 Estimating the Privacy Benefit of Multi-hosted Domains

In this section, we focus on the 81% of multi-hosted domains that can benefit from ESNI, and attempt to assess their actual privacy gain. Recall that a website can gain some privacy benefit only if it is co-hosted with other websites, in which case an on-path adversary will not know which among all co-hosted websites is actually being visited. We use k-anonymity to model and quantify the privacy gain of multi-hosted domains.

Going back to Figure 2, we can apply this definition in two ways, depending on whether we focus on IP addresses or domains. For a given IP address, its k-anonymity value ("k" for brevity) corresponds to the number of co-hosted domains. For a given multi-hosted domain, its k may be different across the individual IP addresses on which it is hosted, as the number of co-hosted domains on each

Table 1: Top hosting providers offering the highest median *k*-anonymity per IP address.

Median *k*	Organization	Unique IPs	Highest Rank
3,311	AS19574 Corporation Service	2	1,471
2,740	AS15095 Dealer Dot Com	1	80,965
2,690	AS40443 CDK Global	1	68,310
1,338	AS32491 Tucows.com	1	22,931
1,284	AS16844 Entrata	1	96,564
946	AS39570 Loopia AB	6	19,238
824	AS54635 Hillenbrand	1	117,251
705	AS53831 Squarespace	23	386
520	AS12008 NeuStar	2	464
516	AS10668 Lee Enterprises	4	3,211

Table 2: Top hosting providers with highest number of observed IP addresses.

Median *k*	Organization	Unique IPs	Highest Rank
16	AS13335 Cloudflare, Inc.	64,285	112
5	AS16509 Amazon.com, Inc.	47,786	37
5	AS46606 Unified Layer	27,524	1,265
3	AS16276 OVH SAS	22,598	621
3	AS24940 Hetzner Online GmbH	21,361	61
4	AS26496 GoDaddy.com, LLC	16,415	90
2	AS14061 DigitalOcean, LLC	11,701	685
3	AS14618 Amazon.com, Inc.	11,008	11
6	AS32475 SingleHop LLC	10,771	174
2	AS26347 New Dream Network	10,657	1,419
7	AS15169 Google LLC	9,048	1
3	AS63949 Linode, LLC	8,062	2,175
4	AS8560 1&1 Internet SE	6,898	2,580
3	AS32244 Liquid Web, L.L.C	6,412	1,681
3	AS19551 Incapsula Inc	6,338	1,072
4	AS36351 SoftLayer Technologies	6,005	483
3	AS16625 Akamai Technologies	5,862	13
4	AS34788 Neue Medien Muennich	5,679	7,526
6	AS9371 SAKURA Internet Inc.	5,647	1,550
3	AS8075 Microsoft Corporation	5,360	20

of those addresses may be different. Consequently, the k value of a multi-hosted domain is calculated as the median k of all its IP addresses.[1] In both cases, the privacy gain increases linearly with k. Based on these definitions, we now explore the privacy gain of domains hosted on different hosting and CDN providers.

Table 1 shows the top-ten hosting providers offering the highest median k-anonymity per IP address (i.e., greater than 500). As shown in the third column, the average number of unique IP addresses observed daily for each provider is very low, with half of them hosting all domains under a single IP address. Using the

[1]Since most domains have similar k values across their hosting IP addresses, both mean and median can be used in this case.

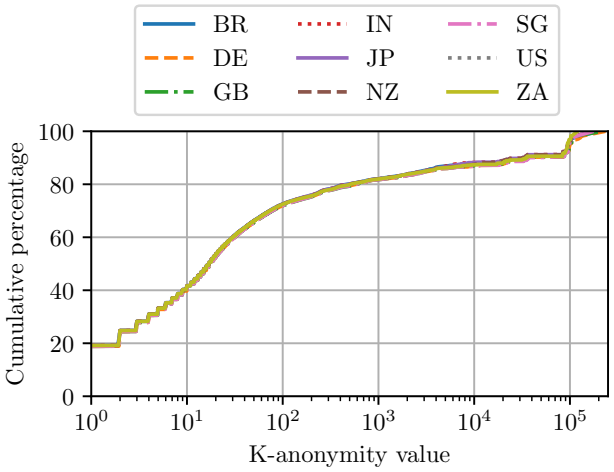

Figure 4: CDF of the *k*-anonymity for all studied domains across nine measurement locations (*k*=1 corresponds to the 18.7% of single-hosted domains).

Hurricane Electric BGP Toolkit, we confirmed that these providers are indeed small, with many of them managing less than 10K IP addresses allocated by regional Internet registries. When looking into the popularity of the websites hosted by these providers, as shown in the last column, the highest ranked website is only at the 386th position, hosted on Squarespace, while more than half of these providers host websites that are well below the top 10K.

Next, we investigate the k-anonymity offered by major providers that dominate the unique IP addresses observed. Table 2 lists the top-20 major hosting and CDN providers with more than 5K unique IP addresses observed. Unlike small hosting providers, these major providers are home to more popular sites. Indeed, the most popular sites hosted by these providers are all within the top 10K. In contrast to small providers, however, the median k-anonymity per IP address offered by these providers is quite low, meaning that sites hosted on them will gain a much lower level of privacy. Except from Cloudflare, which has the highest k of 16, all other providers have a single-digit k.

Tables 1 and 2 represent two ends of the privacy spectrum for multi-hosted domains. On one end, numerous but less popular domains are hosted on providers managing a handful of IP addresses, benefiting from high k-anonymity; on the other end, fewer but more popular websites are hosted on providers managing a much larger pool of millions of IP addresses, suffering from low k-anonymity.

To provide an overall view of the whole privacy spectrum, Figure 4 shows CDFs of k of all studied domains across nine different regions. As illustrated, k values are almost identical across the nine regions from which we conducted our measurements. While our DNS data shows that there are 471K (CDN-supported) domains served from different IP addresses depending on the resolution location, the k values of these domains remain similar regardless of the DNS resolution origin.

As discussed in §3, a low (e.g., single-digit) k cannot allow ESNI to offer meaningful privacy, given that i) not all k sites will be equally popular, and ii) website fingerprinting can be used to improve

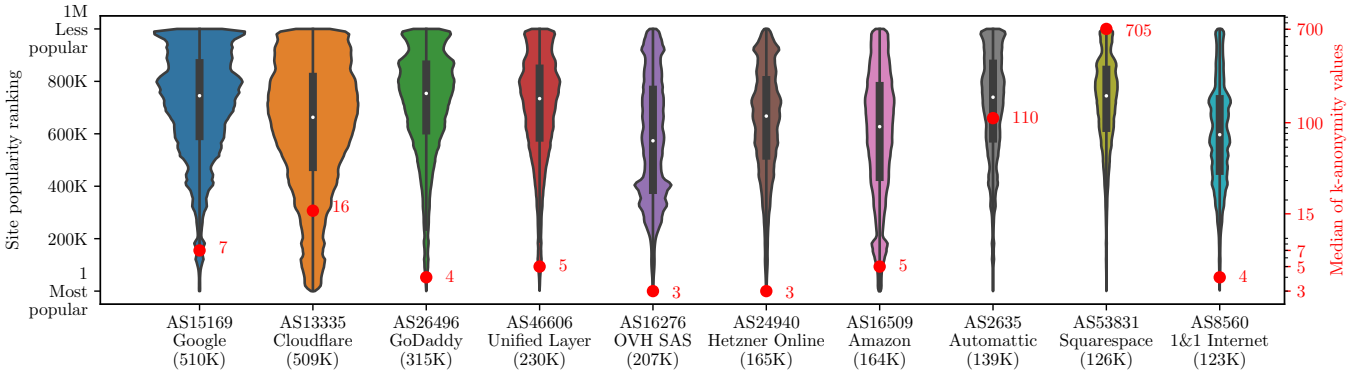

Figure 5: Top providers that host most domains.

attribution accuracy even further [17, 65, 78, 81, 99]. Assuming that k needs to be greater than 100 to provide meaningful privacy, since an adversary would correctly guess a domain being visited with a probability less than 1%, then according to Figure 4, only about 30% of the sites will benefit from domain name encryption. We conduct a more in-depth analysis of the probability with which an adversary would correctly guess domains being visited based on Zipf's law in §5.3.

Finally, we examine the top-10 providers that host the largest number of domains among the ones studied. Although these mostly include some of the providers listed in Table 2, two of them are not included on that table, and one (Squarespace) is actually included in Table 1. The violin plot of Figure 5 depicts the top-ten providers that host most domains. The area of each violin is proportional to the number of domains hosted by that provider, while the shape of each violin illustrates the popularity ranking distribution of hosted websites. The median k of each provider is denoted by the red dot. Google and Cloudflare are the top hosting providers, with more than 500K domains each. Other providers host different numbers of domains, ranging from 315K to 123K.[2] Although hosting fewer domains, both Automattic and Squarespace provide significantly higher privacy with a k of 110 and 705, respectively.

5.3 Weighting the Privacy Benefit Based on Domain Popularity

In §5.2, we used the k-anonymity model to quantify the privacy benefit provided by multi-hosted domains. However, one might consider that the model does not accurately capture a real-world adversary, as not all co-hosted domains are equally popular. Adversaries could base their guess on the probability that a domain is more (or less) likely to be visited, according to the visit frequency of that domain compared to other co-hosted domains. However, it is infeasible for us to obtain the data of domain visit frequencies, since this is only known by the respective hosting providers.

Fortunately, prior studies have shown that the relative visit frequency of domains follows Zipf's law [15, 106]. More specifically, Zipf's law states that the relative probability of a domain (d) being visited is inversely proportional to its popularity ranking

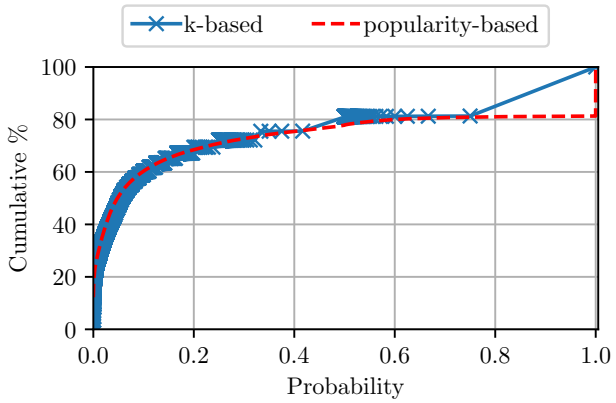

Figure 6: CDF of the probability of correctly guessing a visited domain based on the k-anonymity value and popularity ranking information, as percentage of all tested domains.

($P_d \propto 1/(rank_d)^\alpha$). We thus apply Zipf's law[3] on the popularity ranking of domains to compute the probability with which an adversary can correctly guess that a given domain is being visited.

From a privacy-detrimental IP address, it is straightforward for the adversary to learn the domain being visited as the IP address solely hosts that single domain. However, given a privacy-beneficial IP address that hosts multiple domains, a more realistic adversary would make his guess based on the probability that a domain is more likely to be visited compared to other co-hosted domains. In order to compute this probability, we first obtain the domains $d_1, \ldots, d_n$ that are hosted on a single IP_j and compute their P_d values according to Zipf's law. We define $P_{d_{ij}} = \frac{P_{d_i}}{\sum_{k=1}^{n} P_{d_k}}$ as the probability that domain d_i was visited when IP_j was observed.

For domains that are hosted on multiple IP addresses, the probability is estimated by taking the median of all probabilities that the domain is visited from all IP addresses hosting it. We therefore compute the probability that an adversary can correctly guess domain

[2]Note that a website may be hosted on more than one provider [52]. In that case, we count the site separately for each hosting provider.

[3]For simplicity, we present results with $\alpha = 1$, following the strict Zipf's law. However, adjusting the value of α to match previous observations [15] also gave similar results.

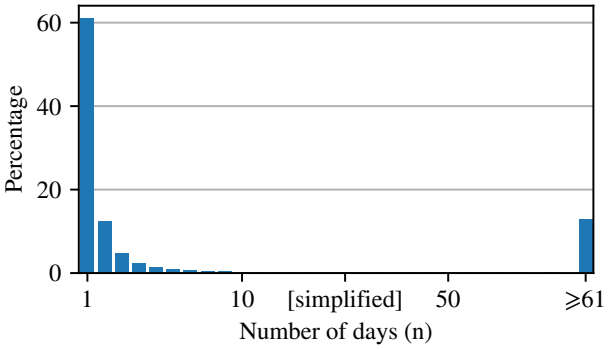

Figure 7: Longevity distribution of domain-to-IP mappings as percentage of number of mappings.

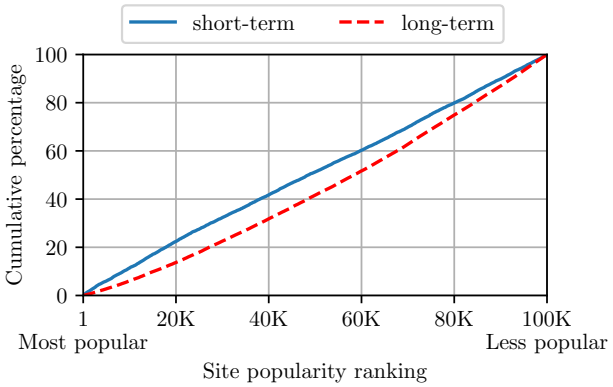

Figure 8: CDF of domain popularity for short-term and long-term domain-to-IP mappings.

d_i that is hosted on $IP_1, \ldots, IP_m$ as follows:

$$P_i = median(P_{d_{i1}}, \ldots, P_{d_{im}}) \quad (1)$$

As shown in Figure 6, our k-anonymity model is a close lower bound to the case where the adversary considers the popularity rankings. The figure shows two CDFs of the probability that the adversary can guess which domain is being visited. The continuous (blue) line is computed based on the k-anonymity value of co-hosted domains. Each domain has an equal probability of $1/k$ to be visited. The dashed (red) line is computed by applying the Zipf's law on the domain popularity. We can see that even if adversaries rely on domain popularity rankings to improve the accuracy of their prediction, the highest probability that this guess is correct is similar to the probability estimated by the k-anonymity value.

5.4 Domain-to-IP Mapping Stability

Besides the degree of co-hosting, the stability of a website's IP address(es) also plays an important role in whether ESNI will provide meaningful privacy benefits. If the IP address of a website changes quite frequently, this will have a positive impact on the privacy offered due to ESNI. Unless adversaries have enough resources to acquire all domain-to-IP mappings of interest at almost real-time, they will no longer be able to use the destination IP address as an accurate predictor of the visited website, because a previously known domain-to-IP mapping may not be valid anymore. On the other hand, mappings that remain stable over the time make it easier for adversaries to monitor the visited websites.

In this section, we examine the stability of domain-to-IP mappings, and how it affects privacy. We are particularly interested in finding how often domain-to-IP mappings change. As discussed in §4, all top lists of popular sites have their own churn (i.e., new sites appear and old sites disappear from the lists on a daily basis). To prevent this churn from affecting our analysis, we consider only the subset of domains that were present daily on both top lists (§4.2) during the whole period of 61 days of our study. This set of domains comprises 2.6M domains, from which we observed a total of 22.7M unique domain-to-IP mappings because a domain may be hosted on hundreds of IP addresses.

Figure 7 shows the distribution of the longevity of these mappings in days. More than 80% of the mappings last less than four consecutive days (*short-term mappings*), corresponding to 202K (7.7%) domains served from 400K unique IP addresses. On the other hand, 13% of the mappings remain unchanged for the whole study period (*long-term mappings*), corresponding to 2.4M domains served from 1.1M unique IP addresses. As also shown in Figure 7, there are two dominant clusters of domains that either change their hosting IP addresses frequently or do not change at all. This is a favorable result for adversaries, as it implies that they do not have to keep resolving a large number of domains, since most domain-to-IP mappings remain quite stable over a long period.

The popularity distribution of the domains that correspond to these two short-term and long-term mappings is shown in Figure 8. While domains with short-term mappings are evenly distributed across the popularity spectrum, domains with long-term mappings slightly lean towards lower popularity rankings. This result can be attributed to the fact that more popular websites are more likely to rotate their IP addresses for load-balancing reasons, while less popular sites are more likely to be served from static IP addresses.

An increased churn of IP addresses also helps ESNI provide better privacy. We thus investigated which providers exhibit the highest churn rate by grouping the IP addresses of short-term mappings according to their ASN. Figure 9 shows the top-ten providers with the highest number of IP addresses in short-term mappings (bars). The dots indicate the number of domains hosted on those IP addresses. Although Amazon and Akamai do not top the list of providers that host most domains (Figure 5), along with Cloudflare they occupy the top five positions of the providers with the highest number of dynamic IPs. Google uses a relatively small pool of around 5.3K IP addresses, to host more domains (41K) than the other providers.

6 COMPARISON WITH OTHER DATASETS

In this section, we analyze existing public DNS datasets to examine the impact of i) larger datasets (in terms of number of domains), and ii) more localized vantage points, on the estimation of per-domain k-anonymity.

The Active DNS Project [61] is currently collecting A records of about 300M domains derived from 1.3K zone files on a daily basis. In addition to this effort, Rapid7 [87] also conducts active DNS measurements at a large scale and offers researchers access to its

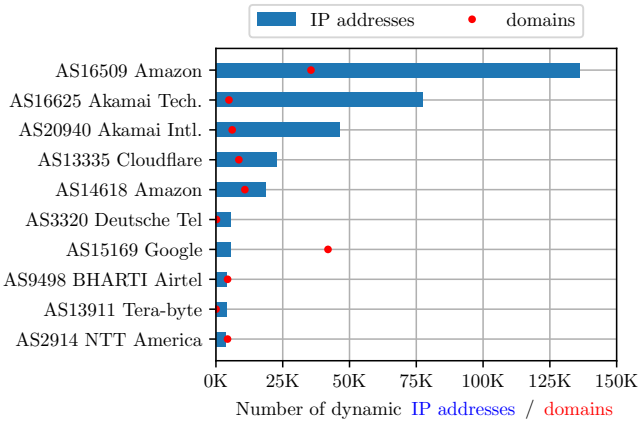

Figure 9: Top providers with the highest number of high-churn IP addresses.

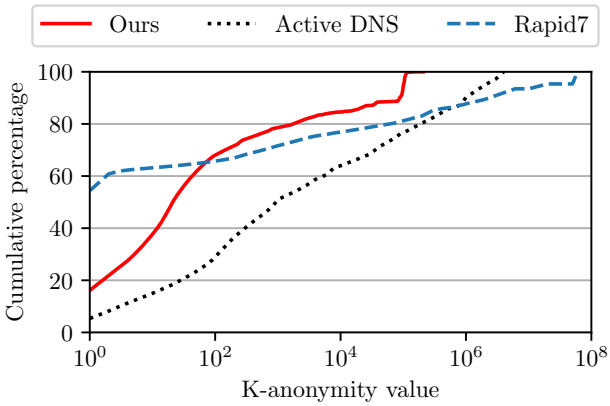

Figure 10: CDF of the *k*-anonymity value per domain as a percentage of *all domains* observed from the Active DNS Project, Rapid7, and our datasets.

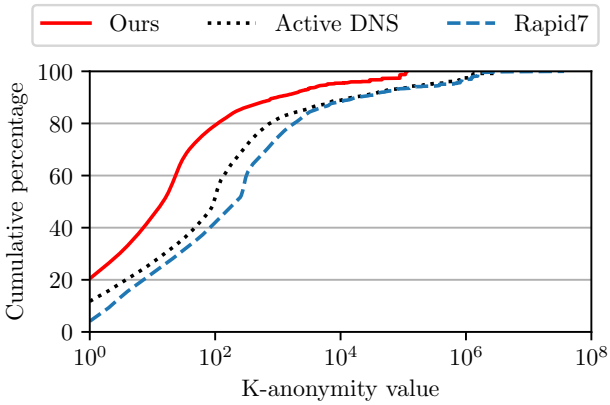

Figure 11: CDF of the *k*-anonymity value per domain as a percentage of *common domains* observed from the Active DNS Project, Rapid7, and our datasets.

data. Unlike the Active DNS Project, Rapid7 resolves a much larger number of domains (1.8B), but with a lower frequency (domains are resolved only once a week). The dataset includes not only second-level domains from several TLD zone files, but also lower-level domains obtained through web crawling and targeted scanning with Zmap [34].

Different from these datasets, as discussed in §4.2, our domain name dataset is curated from the global lists of Alexa and Majestic. We also perform measurements from vantage points around the world to observe localized DNS responses from CDNs. In contrast, the above datasets are collected from local vantage points, as their goal is to maximize the number of observed domains, and not to exhaustively resolve all potential IP addresses of a domain. In particular, the Active DNS Project is run from Georgia Tech, while Rapid7's data is collected using AWS EC2 nodes in the US. To that end, we used two datasets from the Active DNS Project and Rapid7 collected on March 29th, 2019 for our comparison. We sanitized poisoned records from these datasets, as described in Appendix §B, before analyzing them.

Figure 10 shows the CDF of the *k*-anonymity value per domain for all the domains in the Active DNS, Rapid7, and our datasets, while Figure 11 shows the CDF of the *k*-anonymity value per domain for only common domains among the three datasets.

In Figure 10, when $k = 1$, there is some difference in the percentage of single-hosted domains between the Active DNS Project (5.3%), Rapid7 (54.3%), and our observation (18.7%). As expected, the percentage of single-hosted domains for Rapid7 is the highest because this dataset is the largest (1.8B FQDNs), and includes lower-level FQDNs that may host other services (e.g., email, DNS, SSH) instead of web content. On the other hand, only 5.3% of the domains in the Active DNS dataset are single-hosted, since the dataset contains mostly A records of domains extracted from TLD zone files, instead of many lower-level FQDNs.

When all domains are considered, our observation of single-hosted domains is in between the two (18.7%) because (as mentioned in §4.2) we derive our domain name dataset from the two global top websites lists (Alexa and Majestic), and include not only second-level domains but also lower-level FQDNs, as long as they are included in the lists and serve web content. This aligns well with the results of Rapid7 and Active DNS, given the inherent over-approximation of the Rapid7 dataset (due to non-web services and highly unpopular or even single-use domains) and the inherent under-approximation of the Active DNS Project dataset (due to only domains extracted from TLD zone files).

When considering only common domains, the percentages of single-hosted domains are 4.1%, 11.7%, and 20.4% for the Rapid7, Active DNS, and our dataset, respectively. In other words, some domains classified as single-hosted in our dataset are actually multi-hosted when considering the larger datasets. This result confirms our hypothesis discussed in §4.2 that by only considering the two sets of popular websites, we would have missed those less popular, random, or even dormant domains in the long tail. Thus, the result provided by our dataset can be considered as a lower bound value of the actual *k*-anonymity, since any increase in *k* provided by less well-known, random, or even dormant domains is less meaningful from the perspective of adversaries whose goal is to reveal a visited

website by incorporating the popularity ranking information in their "guess."

At the right end of the CDFs in both Figures 10 and 11, the two larger datasets exhibit k-anonymity values higher than ours. The primary reason for this is that these datasets include not only second-level domains, but also third-level and longer FQDNs. The higher k values also comprise the long-tail of domains that are not included in our dataset, including less popular domains, random or single-use FQDNs used for tracking, and malicious domains [96].

7 DISCUSSION

7.1 Recommendations

While the *security* benefits of DoH/DoT against on-path adversaries are clear (e.g., prevention of MiTM or MoTS DNS poisoning attacks), our findings show that Encrypted SNI alone cannot fully address the *privacy* concerns it aims to tackle. More effort and collaboration from all involved parties (i.e., operators of DNS authoritative name servers, website owners, and hosting and CDN providers) are needed. In this section we provide some suggestions for maximizing the privacy benefits of ESNI.

Full Domain Name Confidentiality. In the current designs, plaintext domain names are exposed through two channels: the SNI extension in TLS, and traditional DNS name resolutions—the deployment of DoH/DoT is thus a prerequisite for ESNI. Equivalently, the use of DoH/DoT will not provide any meaningful privacy if domain names are still exposed through the (unencrypted) SNI extension in TLS handshake traffic.

Recently, there is a push for the deployment of DoH/DoT, with major organizations already supporting it (e.g., Google, Cloudflare, Firefox), though this has not been followed by an equivalent effort for the deployment of ESNI. An even more complicated method of securing DNS traffic is DNS-over-HTTPS-over-Tor, which has been already implemented and supported by Cloudflare [91]. Unless the confidentiality of domain names is preserved on both channels (TLS and DNS), neither technology can provide any actual privacy benefit if deployed individually.

Domain Owners. Website owners who want to provide increased privacy to their users can seek hosting providers or CDNs that offer an increased ratio of co-hosted domains per IP address and/or highly dynamic domain-to-IP mappings. In practice, however, this may be challenging. Our results show that unfortunately only a few providers offer a high domain-to-IP ratio, while other more pressing factors (e.g., site popularity) may tilt the decision towards other more important factors, such as latency, bandwidth, or points of presence.

While pointer (PTR) records are often not configured, from a privacy perspective, their operation conflicts with DoH/DoT and ESNI as further discussed in Appendix C. Consequently, website operators should not configure PTR records unless absolutely necessary (e.g., for email servers). In addition, providers with a higher rotation of IP addresses are more preferable, as this also helps in improving privacy.

Hosting Providers. Hosting and CDN providers are in a more privileged position to achieve meaningful impact in helping improve the potential privacy benefits of ESNI, as they can control the number of co-hosted domains per IP address, and the frequency of IP addresses rotation. Unless website owners prefer otherwise, providers could group more websites under the same IP address (which, understandably, may not be desirable for some websites).

To improve k-anonymity even more meaningfully, providers should cluster websites according to similarities in terms of traffic patterns and popularity ranking, to hinder website fingerprinting attempts. As discussed in §5.4, more dynamic hosting IP addresses can also help improve visitors' privacy. Currently, the number of websites benefiting from more short-lived domain-to-IP mappings is relatively small. While more frequent IP address changes may complicate operational issues and are certainly more challenging to deploy from a technical perspective (especially for smaller providers), existing load balancing schemes already provide such a capability, which could be tuned to also maximize privacy. In the future, it may be worthwhile to explore more sophisticated schemes that actively attempt to maximize privacy by increasing the "shuffling" rate of co-hosted domains per IP address, to hinder attribution even further, especially when considering more determined adversaries.

7.2 Impact

The deployment of domain encryption has various advantages and disadvantages from the two—rather conflicting—perspectives of Internet censorship and network visibility.

The existing plaintext exposure of domain names on the wire, as part of DNS requests and TLS handshakes, has enabled the wide use of network traffic filtering and censorship based on domain names [2, 32, 39, 49, 82, 93]. In a future with all domain name information being encrypted, DNS and SNI traffic will no longer be an effective vector to conduct censorship. It is likely that censors will shift to use IP-based blocking, which can be very effective if hosting IP addresses of censored websites are stable and host only a handful of sites or services [33, 50]. However, if providers start adapting according to the above mentioned recommendations, the cost of conducting IP-based blocking will increase, since a censor will have to keep track of which IP address belong to which websites.

More importantly, the collateral damage caused by this type of blocking will also increase dramatically if censored websites are co-hosted with multiple other innocuous websites [52]. Although some previous actions from the side of providers (e.g., hindering domain fronting [40]) have shown that privacy is often given a secondary priority [88], as the collateral damage caused to censors may also impact significantly the providers, the renewed recent focus on privacy as a potential competitive advantage by some providers may encourage the deployment of hosting schemes that will improve the privacy benefits of ESNI.

On the other hand, while providing many security and privacy benefits, domain name encryption can be a "double-edged sword" for network administrators who want to have full visibility and control over domain resolutions in the networks under their responsibility. Until now, the operation of firewalls, intrusion detection systems, and anti-spam or anti-phishing filters has benefited immensely from the domain name information extracted from network traffic, as is evident by the series of works mentioned in §8 that employ DNS data to detect domain name abuses and malicious online activities.

Under a full DoH/DoT and ESNI deployment, this visibility will be lost, and systems based on domain reputation [9] and similar technologies will be severely impacted. While many malicious domains often hide themselves by sharing hosting addresses with other innocuous and unpopular websites [96], it will be challenging to detect and block them. A possible solution would be to rely solely on TLS proxying using custom provisioned certificates, in order to gain back the visibility lost by ESNI and DoH/DoT, which is already a common practice used by transparent SSL/TLS proxies. Although this will defeat any privacy benefits of these technologies, this may be an acceptable trade off for corporate networks and other similar environments.

8 RELATED WORK

The domain name system is one of the core elements of the Internet and plays an essential role for most online services. As a result, it has been (ab)used for many different purposes. In this section, we review prior works that investigate DNS from security and privacy perspectives, and some recent studies that analyze the domain name ecosystem via empirical measurements.

From a security perspective, domain names have been heavily abused for illicit purposes. For instance, domain squatting is one of the most common abuses. It is used to register domains that are similar to those owned by well-known Internet companies. Domain squatting has many variations, including typo-squatting [5, 57, 96], homograph-based squatting [41, 85], homophone-based squatting [76], bit-squatting [77], and combo-squatting [58]. Domain names registered using these squatting techniques can then be used for phishing [80, 85] or distributing malware [8]. To cope with these unwanted domain names, DNS data has been used intensively to create domain name reputation systems to detect abuse [9–11, 62].

Another major form of DNS abuse is DNS poisoning, in which an on-path observer can easily observe and tamper with DNS responses to redirect users to malicious websites or to censor unwanted content [2, 32, 39, 67, 82, 93]. The exposure of domain names in DNS requests and TLS handshakes (due to SNI) has also been extensively used for traffic filtering and censorship [21, 49].

As mentioned in §2.1, the traditional design of DNS exposes Internet users to severe privacy risks. In addition to on-path observers (discussed in §3), previous works have also studied the privacy risk associated with centralizing all domain name resolutions to third-party recursive resolvers (e.g., 8.8.8.8, 1.1.1.1) [18, 19, 46, 59, 95, 105]. Zhao et al. [105] propose to add random noise and use private information retrieval to improve privacy by obfuscating DNS queries. These proposals however have turned out to be impractical and insecure under certain circumstances [19], and have not been adopted. Lu et al. [69] propose privacy-preserving DNS (PPDNS), which is based on distributed hash tables and computational private information retrieval. More recently, Hoang et al. [51] propose K-resolver as a mechanism to distribute DoH queries among multiple recursors, thus exposing to each recursor only a part of a user's browsing history. Sharing similar goals with our study, Shulman et al. [95] examined the pitfalls of DNS encryption. By analyzing the co-residence of zone files on name servers, the authors argue that guessing visited domains by destination IP address does not provide a significant advantage. Our findings, however, show that this is only the case for a small number of domains that are co-hosted with an adequate number of other domains.

Hounsel et al. [54] study the effect of DoH/DoT on performance of domain name resolution and content delivery. The study finds that the resolution time of DoH/DoT is longer than traditional DNS resolution. Of the two new technologies, DoT provides better page load times while DoH at best has the same page load times as DNS. They also find that DoT and DoH perform worse than DNS in networks with sub-optimal performance. Similarly, a recent study by Bottger et al. [14] analyzes the DoH ecosystem and shows that they can obtain more advanced privacy features of DoH with marginal performance degradation in terms of page load times.

There have also been studies that investigated the robustness of the DNS ecosystem through various types of measurements. Ramasubramanian et al. [86] leverage a dataset of almost 600K domains to study their trusted computing base, which is the set of name servers on which a FQDN is hosted. The study shows that a typical FQDN depends on 46 servers on average. Dell'Amico et al. [30] use DNS data collected through both active and passive measurements to also investigate the ecosystem of dependencies between websites and other Internet services. Similarly to our work, Shue et al. [94] use DNS data collected by both passive and active measurements to study web server co-location and shared DNS infrastructure. However, their measurements were conducted from a single location, while excluding all servers belonging to CDNs. Furthermore, the passive DNS dataset used was collected by capturing network traffic from the authors' institute, therefore facing all potential issues of a passive measurement discussed in §4.1. More recently, Hoang et al. [52] revisit the results of Shue et al. [94] by conducting a large-scale active DNS measurement study, which reveals that the Web is still centralized to a handful of hosting providers, while IP blocklists cause less collateral damage than previously observed regardless of a high level of website co-location.

9 CONCLUSION

The deployment of encrypted SNI in TLS, combined with DNS over HTTPS/TLS, will definitely provide many security benefits to Internet users. However, as we have shown in this work, a significant effort is still needed in order for these same technologies to provide meaningful privacy benefits. More specifically, while domain name information is encrypted, the IP address information is still visible to any on-path observers and can be used to infer the websites being visited.

Using DNS data collected through active DNS measurements, we studied the degree of co-hosting of the current web, and its implications in relation to ESNI's privacy benefits. Quantifying these benefits for co-hosted websites using k-anonymity, we observed that the majority of popular websites (about half of all domains studied) will gain only a small privacy benefit (k<16). Such a small degree of co-hosting is not enough to withstand determined adversaries that may attempt to perform attribution by considering the popularity or even the traffic patterns of the co-hosted websites on an observed destination IP address. Domains that will obtain a more meaningful privacy benefit (k>500) include only vastly less popular websites mostly hosted by smaller providers, while 20% of

the websites, will not gain any benefit at all due to their one-to-one mapping between domain name and hosting IP address.

We hope that our findings will raise awareness about the remaining effort that must be undertaken to ensure a meaningful privacy benefit from the deployment of ESNI. In the meantime, privacy-conscious website owners may seek hosting services offered by providers that exhibit a high ratio of co-hosted domains per IP address, and highly dynamic domain-to-IP mappings.

ACKNOWLEDGEMENTS

We are grateful 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 would like to thank all the anonymous reviewers for their thorough feedback on earlier drafts of this paper. We also thank Hyungjoon Koo, Shachee Mishra, Tapti Palit, Seyedhamed Ghavamnia, Jarin Firose Moon, Christine Utz, Shinyoung Cho, Rachee Singh, Thang Bui, and others who preferred to remain anonymous for helpful comments and suggestions.

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] 2012. The Collateral Damage of Internet Censorship by DNS Injection. *SIGCOMM Computer Communications Review* 42, 3 (2012), 21–27. http://www.sigcomm.org/node/3275

[2] 2014. Towards a Comprehensive Picture of the Great Firewall's DNS Censorship. In *4th USENIX Workshop on Free and Open Communications on the Internet (FOCI 14)*. USENIX Association, San Diego, CA.

[3] Josh Aas, Richard Barnes, Benton Case, Zakir Durumeric, Peter Eckersley, Alan Flores-López, J. Alex Halderman, Jacob Hoffman-Andrews, James Kasten, Eric Rescorla, and et al. 2019. Let's Encrypt: An Automated Certificate Authority to Encrypt the Entire Web. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. Association for Computing Machinery, New York, NY, USA, 2473–2487. https://doi.org/10.1145/3319535.3363192

[4] Josh Aas and Sarah Gran. 2019. Let's Encrypt Has Issued a Billion Certificates. https://letsencrypt.org/2020/02/27/one-billion-certs.html

[5] Pieter Agten, Wouter Joosen, Frank Piessens, and Nick Nikiforakis. 2015. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In *Proc. Network and Distributed System Security Symposium (NDSS)*.

[6] Alexa Internet, Inc. Accessed 2019. How are Alexas's traffic rankings determined? https://support.alexa.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined-

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

[8] 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.

[9] Manos Antonakakis, Roberto Perdisci, David Dagon, Wenke Lee, and Nick Feamster. 2010. Building a Dynamic Reputation System for DNS. In *the 19th USENIX Conference on Security*. USENIX Association, Berkeley, CA, USA, 18–18.

[10] Manos Antonakakis, Roberto Perdisci, Wenke Lee, Nikolaos Vasiloglou, II, and David Dagon. 2011. Detecting Malware Domains at the Upper DNS Hierarchy. In *the 20th USENIX Conference on Security*. USENIX Association, Berkeley, CA, USA, 27–27.

[11] Manos Antonakakis, Roberto Perdisci, Yacin Nadji, Nikolaos Vasiloglou, Saeed Abu-Nimeh, Wenke Lee, and David Dagon. 2012. From Throw-away Traffic to Bots: Detecting the Rise of DGA-based Malware. In *the 21st USENIX Conference on Security Symposium*. USENIX Association, Berkeley, CA, USA, 24–24.

[12] Simurgh Aryan, Homa Aryan, and J. Alex Halderman. 2013. Internet Censorship in Iran: A First Look. In *Presented as part of the 3rd USENIX Workshop on Free and Open Communications on the Internet*. USENIX, Washington, D.C.

[13] S. Bortzmeyer and S. Huque. 2016. *NXDOMAIN: There Really Is Nothing Underneath*. RFC 8020. IETF. https://tools.ietf.org/html/rfc8020

[14] Timm Böttger, Felix Cuadrado, Gianni Antichi, Eder Leão Fernandes, Gareth Tyson, Ignacio Castro, and Steve Uhlig. 2019. An Empirical Study of the Cost of DNS-over-HTTPS. In *Proceedings of the Internet Measurement Conference (IMC '19)*. Association for Computing Machinery, New York, NY, USA, 15–21. https://doi.org/10.1145/3355369.3355575

[15] L. Breslau, Pei Cao, Li Fan, G. Phillips, and S. Shenker. 1999. Web caching and Zipf-like distributions: evidence and implications. In *The IEEE Conference on Computer Communications*, Vol. 1. 126–134 vol.1.

[16] Xiang Cai, Rishab Nithyanand, Tao Wang, Rob Johnson, and Ian Goldberg. 2014. A Systematic Approach to Developing and Evaluating Website Fingerprinting Defenses. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. Association for Computing Machinery, New York, NY, USA, 227–238. https://doi.org/10.1145/2660267.2660362

[17] Xiang Cai, Xin Cheng Zhang, Brijesh Joshi, and Rob Johnson. 2012. Touching from a Distance: Website Fingerprinting Attacks and Defenses. In *Proceedings of the ACM Conference on Computer and Communications Security*.

[18] Sergio Castillo-Perez and Joaquin Garcia-Alfaro. 2008. Anonymous Resolution of DNS Queries. In *On the Move to Meaningful Internet Systems: OTM 2008*. Springer Berlin Heidelberg, Berlin, Heidelberg, 987–1000.

[19] S. Castillo-Perez and J. Garcia-Alfaro. 2009. Evaluation of Two Privacy-Preserving Protocols for the DNS. In *2009 Sixth International Conference on Information Technology: New Generations*. 411–416. https://doi.org/10.1109/ITNG.2009.195

[20] Abdelberi Chaabane, Terence Chen, Mathieu Cunche, Emiliano De Cristofaro, Arik Friedman, and Mohamed Ali Kaafar. 2014. Censorship in the wild: Analyzing Internet filtering in Syria. In *Internet Measurement Conference*. ACM, 285–298.

[21] Zimo Chai, Amirhossein Ghafari, and Amir Houmansadr. 2019. On the Importance of Encrypted-SNI (ESNI) to Censorship Circumvention. In *9th USENIX Workshop on Free and Open Communications on the Internet (FOCI 19)*. USENIX Association, Santa Clara, CA.

[22] Taejoong Chung, David Choffnes, and Alan Mislove. 2016. Tunneling for Transparency: A Large-Scale Analysis of End-to-End Violations in the Internet. In *Proceedings of the 2016 Internet Measurement Conference (IMC '16)*. ACM, New York, NY, USA, 199–213. https://doi.org/10.1145/2987443.2987455

[23] Taejoong Chung, Roland van Rijswijk-Deij, Balakrishnan Chandrasekaran, David Choffnes, Dave Levin, Bruce M. Maggs, Alan Mislove, and Christo Wilson. 2017. A Longitudinal, End-to-End View of the DNSSEC Ecosystem. In *26th USENIX Security Symposium*. USENIX Association, Vancouver, BC, 1307–1322.

[24] Richard Clayton, Steven J. Murdoch, and Robert N. M. Watson. 2006. Ignoring the Great Firewall of China. In *Privacy Enhancing Technologies (Lecture Notes in Computer Science)*, Vol. 4258. Springer, Berlin, Heidelberg, 20–35.

[25] Lorenzo Colitti, Steinar H. Gunderson, Erik Kline, and Tiziana Refice. 2010. Evaluating IPv6 Adoption in the Internet. In *Passive and Active Measurement*. Springer Berlin Heidelberg, Berlin, Heidelberg, 141–150.

[26] Jedidiah R. Crandall, Daniel Zinn, Michael Byrd, Earl Barr, and Rich East. 2007. ConceptDoppler: A Weather Tracker for Internet Censorship. In *Computer and Communications Security*. ACM, New York, 352–365.

[27] Weiqi Cui, Tao Chen, Christian Fields, Julianna Chen, Anthony Sierra, and Eric Chan-Tin. 2019. Revisiting Assumptions for Website Fingerprinting Attacks. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security (Asia CCS '19)*. Association for Computing Machinery, New York, NY, USA, 328–339. https://doi.org/10.1145/3321705.3329802

[28] Jakub Czyz, Mark Allman, Jing Zhang, Scott Iekel-Johnson, Eric Osterweil, and Michael Bailey. 2014. Measuring IPv6 Adoption. In *Proceedings of the 2014 ACM Conference on SIGCOMM (SIGCOMM '14)*. ACM, New York, NY, USA, 87–98.

[29] Tianxiang Dai, Haya Shulman, and Michael Waidner. 2016. DNSSEC Misconfigurations in Popular Domains. In *Cryptology and Network Security*. Springer, 651–660.

[30] 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.

[31] T. Dierks and C. Allen. 1999. *The TLS Protocol Version 1.0*. RFC 2246. IETF. https://tools.ietf.org/html/rfc2246

[32] 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*.

[33] 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. https://www.usenix.org/conference/foci18/presentation/dunna

[34] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. 2013. ZMap: Fast Internet-wide Scanning and Its Security Applications. In *Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13)*. USENIX, Washington, D.C., 605–620. https://www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/durumeric

[35] K. P. Dyer, S. E. Coull, T. Ristenpart, and T. Shrimpton. 2012. Peek-a-Boo, I Still See You: Why Efficient Traffic Analysis Countermeasures Fail. In *Proceedings of the IEEE Symposium on Security & Privacy*.

[36] D. Eastlake and C. Kaufman. 1997. *Domain Name System Security Extensions*. RFC 2065. IETF. https://tools.ietf.org/html/rfc2065

[37] H. Eidnes, G. de Groot, and P. Vixie. 1998. *Classless IN-ADDR.ARPA delegation*. RFC 2317. IETF. https://www.ietf.org/rfc/rfc2317

[38] Let's Encrypt. 2019. Let's Encrypt Stats. https://letsencrypt.org/stats/

[39] 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.

[40] 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.

[41] Evgeniy Gabrilovich and Alex Gontmakher. 2002. The Homograph Attack. *Commun. ACM* 45, 2 (Feb. 2002), 128–. https://doi.org/10.1145/503124.503156

[42] Google. 2018. DNS-over-HTTPS. https://developers.google.com/speed/public-dns/docs/dns-over-https Accessed: October 2018.

[43] Google. 2019. DNS-over-TLS. https://developers.google.com/speed/public-dns/docs/dns-over-tls Accessed: March 2019.

[44] Shuai Hao, Yubao Zhang, Haining Wang, and Angelos Stavrou. 2018. End-Users Get Maneuvered: Empirical Analysis of Redirection Hijacking in Content Delivery Networks. In *27th USENIX Security Symposium*. USENIX Association, Baltimore, MD, 1129–1145.

[45] Jamie Hayes and George Danezis. 2016. k-fingerprinting: A Robust Scalable Website Fingerprinting Technique. In *25th USENIX Security Symposium (USENIX Security 16)*. USENIX Association, Austin, TX, 1187–1203. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/hayes

[46] Dominik Herrmann, Karl-Peter Fuchs, Jens Lindemann, and Hannes Federrath. 2014. EncDNS: A Lightweight Privacy-Preserving Name Resolution Service. In *Computer Security - ESORICS 2014*. Springer, 37–55.

[47] Dominik Herrmann, Rolf Wendolsky, and Hannes Federrath. 2009. Website Fingerprinting: Attacking Popular Privacy Enhancing Technologies with the Multinomial Naïve-Bayes Classifier. In *Proceedings of the 2009 ACM Workshop on Cloud Computing Security (CCSW '09)*. Association for Computing Machinery, New York, NY, USA, 31–42. https://doi.org/10.1145/1655008.1655013

[48] Nguyen Phong Hoang, Yasuhito Asano, and Masatoshi Yoshikawa. 2016. Your Neighbors Are My Spies: Location and other Privacy Concerns in GLBT-focused Location-based Dating Applications. *Transactions on Advanced Communications Technology (TACT)* 5, 3 (May 2016), 851–860. https://doi.org/10.23919/ICACT.2017.7890236

[49] 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.

[50] 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.

[51] Nguyen Phong Hoang, Ivan Lin, Seyedhamed Ghavamnia, and Michalis Polychronakis. 2020. K-resolver: Towards Decentralizing Encrypted DNS Resolution. In *Proceedings of The NDSS Workshop on Measurements, Attacks, and Defenses for the Web 2020 (MADWeb '20)*. Internet Society, 7.

[52] Nguyen Phong Hoang, Arian Akhavan Niaki, Michalis Polychronakis, and Phillipa Gill. 2020. The Web is Still Small After More Than a Decade: A Revisit Study of Web Co-location. *SIGCOMM Comput. Commun. Rev.* (2020).

[53] P. Hoffman and P. McManus. 2018. *DNS Queries over HTTPS (DoH)*. RFC 8484. IETF. https://tools.ietf.org/html/rfc8484

[54] Austin Hounsel, Kevin Borgolte, Paul Schmitt, Jordan Holland, and Nick Feamster. 2019. Analyzing the Costs (and Benefits) of DNS, DoT, and DoH for the Modern Web. In *Proceedings of the Applied Networking Research Workshop (ANRW '19)*. ACM, New York, NY, USA, 20–22. https://doi.org/10.1145/3340301.3341129

[55] Z. Hu, L. Zhu, J. Heidemann, A. Mankin, D. Wessels, and P. Hoffman. 2016. *Specification for DNS over Transport Layer Security (TLS)*. RFC 7858. IETF. https://tools.ietf.org/html/rfc7858

[56] Huawei. 2011. *Transport Layer Security (TLS) Extensions: Server Name Indication*. RFC 6066. IETF. https://tools.ietf.org/html/rfc6066#section-3

[57] M. T. Khan, X. Huo, Z. Li, and C. Kanich. 2015. Every Second Counts: Quantifying the Negative Externalities of Cybercrime via Typosquatting. In *2015 IEEE Symposium on Security and Privacy*. 135–150. https://doi.org/10.1109/SP.2015.16

[58] Panagiotis Kintis, Najmeh Miramirkhani, Charles Lever, Yizheng Chen, Rosa Romero-Gómez, Nikolaos Pitropakis, Nick Nikiforakis, and Manos Antonakakis. 2017. Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse. In *the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, USA, 569–586. https://doi.org/10.1145/3133956.3134002

[59] Panagiotis Kintis, Yacin Nadji, David Dagon, Michael Farrell, and Manos Antonakakis. 2016. Understanding the Privacy Implications of ECS. In *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 343–353.

[60] Platon Kotzias, Abbas Razaghpanah, Johanna Amann, Kenneth G. Paterson, Narseo Vallina-Rodriguez, and Juan Caballero. 2018. Coming of Age: A Longitudinal Study of TLS Deployment. In *Proceedings of the Internet Measurement Conference 2018*. ACM, New York, NY, USA, 415–428.

[61] 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*. Springer, 188–208.

[62] Srinivas Krishnan and Fabian Monrose. 2011. An Empirical Study of the Performance, Security and Privacy Implications of Domain Name Prefetching. In *Proceedings of the 2011 IEEE/IFIP 41st International Conference on Dependable Systems&Networks (DSN '11)*. IEEE Computer Society, Washington, DC, USA, 61–72. https://doi.org/10.1109/DSN.2011.5958207

[63] 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.

[64] 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

[65] Marc Liberatore and Brian Neil Levine. 2006. Inferring the Source of Encrypted HTTP Connections. In *Proceedings of the 13th ACM Conference on Computer and Communications Security*.

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

[67] Graham Lowe, Patrick Winters, and Michael L Marcus. 2007. The great DNS wall of China. (2007).

[68] Liming Lu, Ee-Chien Chang, and Mun Choon Chan. 2010. Website Fingerprinting and Identification Using Ordered Feature Sequences. In *Computer Security – ESORICS 2010*, Dimitris Gritzalis, Bart Preneel, and Marianthi Theoharidou (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 199–214.

[69] Y. Lu and G. Tsudik. 2010. Towards Plugging Privacy Leaks in the Domain Name System. In *2010 IEEE Tenth International Conference on Peer-to-Peer Computing (P2P)*. 1–10. https://doi.org/10.1109/P2P.2010.5569976

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

[71] MaxMind. 2019. MaxMind GeoLite2 Databases. https://www.maxmind.com/

[72] Patrick McManus. 2018. Improving DNS Privacy in Firefox. https://blog.nightly.mozilla.org/2018/06/01/improving-dns-privacy-in-firefox/.

[73] Zubair Nabi. 2013. The Anatomy of Web Censorship in Pakistan. In *FOCI*. USENIX, Berkeley, CA, Article 2, 7 pages.

[74] Milad Nasr, Amir Houmansadr, and Arya Mazumdar. 2017. Compressive Traffic Analysis: A New Paradigm for Scalable Traffic Analysis. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*. Association for Computing Machinery, New York, NY, USA, 2053–2069. https://doi.org/10.1145/3133956.3134074

[75] 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*.

[76] Nick Nikiforakis, Marco Balduzzi, Lieven Desmet, Frank Piessens, and Wouter Joosen. 2014. Soundsquatting: Uncovering the Use of Homophones in Domain Squatting. In *Information Security*. Springer, 291–308.

[77] Nick Nikiforakis, Steven Van Acker, Wannes Meert, Lieven Desmet, Frank Piessens, and Wouter Joosen. 2013. Bitsquatting: Exploiting Bit-flips for Fun, or Profit?. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 989–998. https://doi.org/10.1145/2488388.2488474

[78] Andriy Panchenko, Fabian Lanze, Jan Pennekamp, Thomas Engel, Andreas Zinnen, Martin Henze, and Klaus Wehrle. 2016. Website Fingerprinting at Internet Scale. In *Proceedings of NDSS '16*.

[79] Andriy Panchenko, Lukas Niessen, Andreas Zinnen, and Thomas Engel. 2011. Website Fingerprinting in Onion Routing Based Anonymization Networks. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society (WPES '11)*. Association for Computing Machinery, New York, NY, USA, 103–114. https://doi.org/10.1145/2046556.2046570

[80] 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

[81] Simran Patil and Nikita Borisov. 2019. What Can You Learn from an IP?. In *Proceedings of the Applied Networking Research Workshop (ANRW '19)*. ACM, New York, NY, USA, 45–51. https://doi.org/10.1145/3340301.3341133

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

Table 3: Breakdown of the five largest TLDs studied.

	Daily	Total
TLDs	1,031	1,125
FQDNs	7,556,066	13,597,409
.com	3,835,080	7,026,005
.org	347,993	584,924
.de	264,057	501,597
.net	263,262	442,729
.ru	210,701	346,194

Security Symposium (USENIX Security 17).
[83] Matthew Prince. 2018. Encrypting SNI: Fixing One of the Core Internet Bugs. https://blog.cloudflare.com/esni/. Online; accessed September 2018.
[84] quantcast. Accessed 2019. Quantcast Top Websites. Web page. https://www.quantcast.com/top-sites/
[85] 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)*.
[86] Venugopalan Ramasubramanian and Emin Gün Sirer. 2005. Perils of Transitive Trust in the Domain Name System. In *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement (IMC '05)*. USENIX Association, Berkeley, CA, USA, 35–35. http://dl.acm.org/citation.cfm?id=1251086.1251121
[87] Rapid7. 2019. Rapid7: Open Data. https://opendata.rapid7.com/.
[88] Fahmida Y. Rashid. 2018. Amazon joins Google in shutting down doamin fronting. https://duo.com/decipher/amazon-joins-google-in-shutting-down-domain-fronting.
[89] E. Rescorla, K. Oku, N. Sullivan, and C. Wood. 2019. *Encrypted Server Name Indication for TLS 1.3*. Internet Draft. IETF. https://tools.ietf.org/html/draft-ietf-tls-esni-03
[90] 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.
[91] Mahrud Sayrafi. 2018. Introducing DNS Resolver for Tor. https://blog.cloudflare.com/welcome-hidden-resolver/. Online; accessed September 2018.
[92] 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
[93] Will Scott, Thomas Anderson, Tadayoshi Kohno, and Arvind Krishnamurthy. 2016. Satellite: Joint Analysis of CDNs and Network-Level Interference. In *USENIX Annual Technical Conference*.
[94] 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
[95] Haya Shulman. 2014. Pretty Bad Privacy: Pitfalls of DNS Encryption. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES '14)*. ACM, New York, NY, USA, 191–200. https://doi.org/10.1145/2665943.2665959
[96] Janos Szurdi, Balazs Kocso, Gabor Cseh, Jonathan Spring, Mark Felegyhazi, and Chris Kanich. 2014. The Long "Taile" of Typosquatting Domain Names. In *23rd USENIX Security Symposium (USENIX Security 14)*. USENIX Association, San Diego, CA, 191–206. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/szurdi
[97] Cisco Umbrella. Accessed 2019. Umbrella Popularity List. Web page. https://s3-us-west-1.amazonaws.com/umbrella-static/index.html
[98] Verisign. 2019. *The Domain Name Industry Brief*. Technical Report. Verisign. https://www.verisign.com/assets/domain-name-report-Q12019.pdf
[99] Tao Wang, Xiang Cai, Rishab Nithyanand, Rob Johnson, and Ian Goldberg. 2014. Effective Attacks and Provable Defenses for Website Fingerprinting. In *Proceedings of the USENIX Security Symposium*.
[100] Florian Weimer. 2005. Passive DNS replication. In *FIRST conference on computer security incident*. 98.
[101] Charles V Wright, Scott E Coull, and Fabian Monrose. 2009. Traffic Morphing: An Efficient Defense Against Statistical Traffic Analysis.. In *NDSS*.
[102] Joss Wright. 2014. Regional Variation in Chinese Internet Filtering. *Information, Communication & Society* 17, 1 (2014), 121–141.
[103] Xueyang Xu, Z. Morley Mao, and J. Alex Halderman. 2011. Internet Censorship in China: Where Does the Filtering Occur?. In *Passive and Active Measurement (LNCS)*, Vol. 6579. Springer, Berlin, Heidelberg, 133–142.
[104] Young Xu. 2016. *Deconstructing the Great Firewall of China*. Technical Report. Thousand Eyes. https://blog.thousandeyes.com/deconstructing-great-firewall-china/
[105] Fangming Zhao, Yoshiaki Hori, and Kouichi Sakurai. 2007. Two-Servers PIR Based DNS Query Scheme with Privacy-Preserving. In *Proceedings of the The 2007 International Conference on Intelligent Pervasive Computing (IPC '07)*. IEEE Computer Society, Washington, DC, USA, 299–302. https://doi.org/10.1109/IPC.2007.107
[106] George Kingsley Zipf. 1929. Relative frequency as a determinant of phonetic change. *Harvard studies in classical philology* 40 (1929), 1–95.
[107] J. Zittrain and B. Edelman. 2003. Internet filtering in China. *IEEE Internet Computing* 7, 2 (March 2003), 70–77. https://doi.org/10.1109/MIC.2003.1189191
[108] Ólafur Guðmundsson. 2018. Introducing DNS Resolver, 1.1.1.1. https://blog.cloudflare.com/dns-resolver-1-1-1-1. Online; accessed September 2018.

A BREAKDOWN OF DOMAINS STUDIED

As shown in Table 3, our derived list comprises an average of 7.5M popular fully qualified domain names (FQDNs) collected on a daily basis, covering 1,031 TLDs. For the whole experiment duration, we studied a total of 13.6M domains and 1,125 TLDs. Table 3 also shows the top five largest TLDs in our dataset, with `.com` being the most dominant, comprising more than 50% of the domains observed.

B POISONED RESPONSE SANITIZATION

While processing public DNS datasets from other sources (to which we compare our findings in §6), we surprisingly discovered thousands of low-ranked or obscure domains seemingly being co-hosted on the same IP addresses that also host very popular websites, such as Facebook and Twitter—which of course was not actually the case. As part of our investigation, we observed that the authoritative servers of most of these domains were located in China (using the MaxMind dataset [71]). We then queried the same domains from outside China using their authoritative servers, and indeed received responses pointing to IP addresses that belong to either Facebook or Twitter. By inspecting network traffic captures taken during these name resolutions, we observed that the initial response containing the wrong (falsified) IP address was followed by another DNS response with the same valid DNS query ID that contained a different (correct) IP address.

We attribute the above observed behavior to DNS-based censorship by the "Great Firewall" (GFW) of China [1, 24, 67, 102, 104, 107], which has also been observed and analyzed by previous studies [2, 39]. Censorship leakage happens due to the GFW's filtering design, which inspects and censors both egress and ingress network traffic. While some censors (e.g., Pakistan, Syria, Iran) forge DNS responses with NXDOMAIN [13, 20, 22, 73] or private addresses [12], making them easier to distinguish, China poisons DNS responses with routable public IP addresses belonging to other non-Chinese organizations [39, 49, 104]. In contrast to the findings of previous works, however, in this case the real hosting IP addresses of the censored domains are located within China, while previous works mostly focus on investigating the blockage of websites that are hosted outside China (e.g., google.com, facebook.com, blogger.com).

To validate our findings, we cross-checked the IP addresses from second (real) DNS responses with the ones obtained by resolving the same domains from locations in China. As the authoritative servers of these domains are also in China, our queries did not cross the GFW, which mostly filters traffic at border ASes [2, 26, 103], and thus were not poisoned.

Table 4: Most frequently abused subnets in poisoned DNS responses from China.

AS32934 Facebook	AS13414 Twitter	AS36351 SoftLayer
31.13.72.0	199.59.148.0	74.86.12.0
31.13.69.0	199.59.149.0	67.228.235.0
31.13.73.0	199.16.156.0	74.86.151.0
31.13.66.0	199.59.150.0	75.126.124.0
69.171.245.0	199.16.158.0	67.228.74.0

We follow this verification technique, where we issue additional queries to resolve domains whose authoritative name server is located in China and then detect injected DNS packets to exclude poisoned responses from analyses. In total, we detected more than 21K domains based in China with poisoned responses. Table 4 shows the top /24 IP subnets belonging to Facebook, Twitter, and SoftLayer, which are the most frequently observed in poisoned responses. Our observation aligns with recent findings of other censorship measurement studies [49, 75].

C REVERSE DNS LOOKUPS

The Internet Engineering Task Force (IETF) recommends that it should be possible to conduct a reverse DNS lookup for every given domain [37]. In a forward DNS lookup, a domain name is resolved to an A (IPv4) record, while a reverse DNS lookup sends out an IP address to ask for its associated FQDN. The DNS record storing this information is called a PTR (pointer) record. Unless configured to point to a FQDN by its owner, it is not compulsory to configure PTR records for every IP address.

Although performing reverse DNS lookups seems to be a straightforward way of mapping a given IP address back to its associated FQDN, this can potentially uncover only single-hosted domains. More importantly, not all reverse DNS queries return a (meaningful[4]) domain name because the IETF's recommendation is only optional, and thus not adopted universally.

We analyzed Rapid7's reverse DNS dataset, which contains PTR records for the whole public IPv4 space, to see how many IP addresses could be used to reveal the visited destinations under the assumed global ESNI deployment, by just performing a reverse DNS lookup. We find that there are 1.27B unique IP addresses having PTR records. Of these, at least 172M (14%) of them point to a meaningful FQDN (i.e., not in the form of dash-separated IP segments). Within this set of domains, we could find 136K single-hosted domains observed by our dataset. This means about 10% of single-hosted domains have PTR records configured. Under a global ESNI deployment, IP addresses of these domains would be detrimental to the privacy of users who connect to them.

As expected, we also observed more than 3M PTR records in which domain names explicitly indicate through the prefix "mail" that they correspond to email servers. These email servers support PTR record because many providers will not accept messages from other mail servers that do not support reverse lookups [66].

[4]Many PTR records are formatted with dash-separated IP segments. For example, the Amazon EC2 IP address *54.69.253.182* has a PTR record to *ec2-54-69-253-182.us-west-2.compute.amazonaws.com*, which may not actually correspond to a user-facing web service.