Towards a Complete View of the Certificate Ecosystem

Benjamin VanderSloot[†] Johanna Amann[‡] Matthew Bernhard[†]
Zakir Durumeric[†‡] Michael Bailey[§] J. Alex Halderman[‡]

[†] University of Michigan [‡] International Computer Science Inst. [§] Univ. of Illinois Urbana-Champaign
{benvds, matber, zakir, jhalderm}@umich.edu, johanna@icir.org, mdbailey@illinois.edu

ABSTRACT

The HTTPS certificate ecosystem has been of great interest to the measurement and security communities. Without any ground truth, researchers have attempted to study this PKI from a variety of fragmented perspectives, including passively monitored networks, scans of the popular domains or the IPv4 address space, search engines such as Censys, and Certificate Transparency (CT) logs. In this work, we comparatively analyze all these perspectives. We find that aggregated CT logs and Censys snapshots have many properties that complement each other, and that together they encompass over 99% of all certificates found by any of these techniques. However, they still miss 1.5% of certificates observed in a crawl of all domains in `.com`, `.net`, and `.org`. We go on to illustrate how this combined perspective affects results from previous studies. In light of these findings, we have worked with the operators of Censys to incorporate CT log data into its results going forward, and we recommend that future HTTPS measurement adopt this new vantage.

1. INTRODUCTION

Nearly all secure web communication takes place over HTTPS. Both the underlying TLS protocol and the supporting certificate public key infrastructure (PKI) have been studied extensively over the past five years, with questions ranging from understanding the behavior of certificate authorities [11, 14] to detecting server-side vulnerabilities and tracking how quickly they are mitigated [7, 10, 28].

Such measurements are difficult to conduct well, since there is no comprehensive set of trusted certificates or of HTTPS websites — no ground truth for studying this ecosystem. Instead, researchers have attempted to gain visibility into it using various fragmentary perspectives — such as scanning the IPv4 address space [12], querying popular Alexa domains, passively monitoring network traffic [3], and querying Certificate Transparency (CT) logs [20, 21]. Each methodology provides an imperfect view of the world, yet there has been little work to analyze how they differ or how they might be combined to piece together a more comprehensive picture.

Consider, for example, the different perspectives provided by CT logs and the Censys search engine [8], two widely used sources of certificate data. CT is designed to enable auditing of trusted certificates by recording them in publicly verifiable logs. While this may someday provide a complete view of the certificate ecosystem, at present, publishing certificates to CT logs is voluntary in most cases. In contrast, Censys provides a public database of certificates collected by actively scanning the IPv4 address space and Alexa Top Million domains. Although IPv4 scanning might seem to promise an exhaustive view of certificates in use on the public Internet, it misses several important cases, including those served exclusively over IPv6. IP-based scanning also cannot provide the TLS Server Name Indication (SNI) header [13], which specifies the requested domain name and is necessary when a server hosts multiple sites from a single IP address.

In this work, we comparatively analyze the certificates seen by eight measurement perspectives: (1) a Censys certificate snapshot, (2) an exhaustive IPv4 scan on TCP/443, (3) a scan of Alexa Top Million domains, (4) a snapshot of public CT Logs, (5) a scan of domains contained in these CT logs, (6) a scan of domains contained in the `.com`, `.net`, and `.org` zone files [30], (7) a scan of domains from the Common Crawl dataset [6], and (8) certificates passively observed by the ICSI SSL Notary using passive network monitoring [3].

Combining these datasets, we observe nearly 17 million unique browser-trusted certificates that were valid during our measurement interval, August 29 to September 8, 2016. Of these, 90.5% appeared in public CT logs and 38.0% were seen by Censys. To understand this difference, we investigate the impact of SNI by attempting connections to 30 million domains extracted from certificates in CT logs. Only 35% of domains that accepted a connection with SNI offered the same certificate when SNI was not used. This places an

IMC 2016 November 14–16, 2016, Santa Monica, CA, USA

ACM ISBN 978-1-4503-4526-2/16/11.

DOI: http://dx.doi.org/10.1145/2987443.2987462

Perspective	Valid Certificates	Exclusive Certificates	FQDNs	Sites	Description
CT Logs	15,374,936	6,830,849	29,967,065	12,202,712	Certificates in the well known CT logs
Censys Snapshot	6,448,588	609,773	14,495,436	4,817,530	Database of IPv4 and Alexa Top 1M scan results
CT Scan	6,419,584	54,225	20,967,351	9,153,162	Scan of all FQDNs in CT logs
IPv4 HTTPS Scan	3,760,360	637	8,209,465	2,926,766	HTTPS scan of all IPv4 addresses
`.com`, `.net`, `.org` zones	2,436,425	31,195	11,892,649	5,669,024	Root and `www.` names in `.com`, `.net`, and `.org`
Common Crawl	1,258,886	1,154	6,177,985	2,885,466	Scan of domains crawled by Common Crawl
Alexa Top 1M	288,670	0	1,907,256	917,259	HTTPS scan of Alexa Top Million domains
ICSI Notary	256,869	3,805	2,040,138	916,612	Observed in passive network traffic inspection
"Universe"	16,989,236	—	32,454,062	12,673,515	Certificates in any of the perspectives above

Table 1: **Certificate Perspectives** — We compare eight distinct perspectives on the universe of valid certificates, spanning six measurement techniques. Certificate Transparency (CT) is the largest perspective, including many certificates only seen in CT.

upper bound on the certificates observable by IP-based scanning. Combining data from Censys and CT covers 99.4% of all trusted certificates seen by any perspective we studied, and may closely approximate the public HTTPS ecosystem. However, as the vast majority of the certificates in our data originate from these two sources, this number is suspect. To better validate the fraction of certificates visible with these perspectives, we consider certificates seen by scanning domains from the `.com`, `.net`, and `.org` zone files, and find that the union of CT logs and Censys contains 98.5% of them.

Based on these results, we recommend that researchers performing future HTTPS measurements use a combination of data published in CT logs and Censys-style IPv4 scanning. To facilitate this, we are working with the operators of Censys to implement synchronization between Censys and CT logs. Going forward, Censys will continuously incorporate certificate data from public CT logs in its results and publish newly discovered certificates back to Google CT logs, making either data source a strong foundation for studying the certificate ecosystem.

2. CERTIFICATE PERSPECTIVES

In order to compare techniques for measuring certificates, we conducted six kinds of scans and analyzed two existing datasets. Table 1 summarizes these perspectives, which we describe in detail below.

Certificate Transparency Logs.

Certificate Transparency (CT) aims to allow public auditing of trusted certificates [21]. Anyone can submit valid certificates to CT log servers, which record them in cryptographically verifiable public ledgers. Although there is no universal requirement for submission, Google records all certificates seen in its web crawls to CT logs. Chrome requires all issuers submit extended validation (EV) certificates to at least two logs [5]. Chrome recently mandated that Symantec certificates signed after June 1, 2016, be submitted to be trusted as well [26]. Several CAs voluntarily log all certificates they issue, notably Let's Encrypt [23] and StartCom [27].

We retrieved the certificates stored in twelve well-known CT logs on September 8, 2016. These logs are operated by Google ("Pilot", "Aviator", "Rocketeer", and "Submariner"), Digicert, StartCom, Izenpe, Symantec, Venafi, WoSign, CNNIC, and Shengnan GDCA.

Censys Certificate Snapshot.

The Censys search engine [8] publishes daily snapshots of all the certificates it indexes. Censys collects certificates by exhaustively scanning the IPv4 address space without SNI, and by connecting to all Alexa Top Million domains with SNI. Our perspective is based on the September 8, 2016 snapshot.

Scan of FQDNs from CT.

We extracted the fully qualified domain names (FQDNs) from all certificates in our CT log snapshots, covering the common name (CN) and subject alternative name (SAN) fields. We then used ZGrab [8] to attempt an HTTPS connection to each domain, with SNI enabled. The scan ran from the University of Michigan on August 29 and September 6 and 8, 2016.

IPv4 HTTPS Scan.

We used the ZMap suite [12] to scan the IPv4 address space for HTTPS servers listening on TCP/443. The scan took place on August 29, 2016, from the University of Michigan. For each listening host, we attempted a TLS handshake and recorded the presented certificate chain. Since these connections were based on IP addresses rather than domain names, they did not include the SNI header.

Authoritative Zone Files.

We attempted HTTPS handshakes with all domains in the authoritative zone file [30] for `.com`, `.net`, and `.org` domains, for both the base domain and the `www` subdomain. (Since the TLD zone files contain only the name server entries for each domain, we learn only the base domain name.) We ran these scans using ZGrab on August 29 and September 2, 2016, from the University of Michigan. There were 153 million unique domains in these zone files, and we completed 42 million successful HTTPS handshakes to the base domains, and 40 million successful HTTPS handshakes to the `www` subdomains. While we connected to many domains, the certificates served were often only valid for the hosting provider's domain name and not the scanned domain.

Common Crawl.

The Common Crawl project [6] aims to perform a regular, complete crawl of public websites. We processed the January 2016 crawl and extracted 28.9 million unique domains. We used ZGrab to attempt an HTTPS connection to every domain on September 3, 2016, from the University of Michigan.

Alexa Top Million HTTPS Scan.

We used the ZMap suite to attempt connections to the Alexa Top Million domains. The scan took place on September 3, 2016, from the University of Michigan. For each listening host, we attempted a TLS handshake with SNI enabled and recorded the presented certificate chain.

ICSI SSL Notary.

The SSL Notary dataset consists of daily Internet traffic from approximately 180,000 users at five North American academic or research institutions [3]. We analyzed 2.2 billion TLS connections on TCP/443 from July 29 to August 29, 2016, extracting a total of 635,314 certificates. We excluded incomplete connections as well as HPC, Grid, and Tor certificates, resulting in 386,051 certificates, of which 256,869 were trusted by the Mozilla NSS root store.

Due to a nondisclosure agreement that limited our internal data sharing, Notary certificates are not included in cases where we consider the union of all perspectives. This reduces the size of the union by 0.02%.

In total across all these perspectives, we discovered 17 million unique certificates that were valid and trusted by the Mozilla NSS root store. Since the different datasets contain somewhat different temporal perspectives, we consider certificates to be valid only if their date ranges cover our entire collection period, August 29 to September 8. By constraining our data in this way, we ensure that no data source contains certificates that would be invalid in another data source due to the time when the certificates were validated.

2.1 Ethical Considerations

For our active scanning, we honored the University of Michigan's institutional blacklist to exclude endpoints that previously requested not to be scanned. We also followed the best practices defined by Durumeric et al. [12]; we refer to that work for more discussion of the ethics of active scanning. Passive data collection was cleared by the responsible parties at each contributing institution. The ICSI SSL Notary stores connection metadata (e.g., certificate and cipher information) without collecting any connection payload.

3. RESULTS

Our certificate "universe" consists of 16,989,236 unique, valid certificates from our eight perspectives. These certificates contain 32,454,062 FQDNs from 12,673,515 sites (as defined by the public suffix list [25]). The CT logs and Censys snapshot are the two largest datasets. The CT logs contain 15,374,936 certificates (90.5% of certificates observed in this study). Censys sees 6,448,588 certificates — 38% coverage of the certificates observed in this study. When combined, these two perspectives provide 99.4% coverage of all certificates we observe and 99.7% coverage of sites.

	Certificates missing from CT	Fraction missing	Fraction of universe
All	1,614,300 (100.0%)	9.5%	9.5%
GoDaddy	314,966 (19.5%)	29.6%	2.0%
cPanel	120,907 (7.5%)	28.4%	0.7%
Thawte	56,078 (3.5%)	18.0%	0.3%
Starfield	38,220 (2.4%)	34.5%	0.2%
Other	1,084,129 (67.2%)	7.2%	6.4%
`mail`	188,109 (11.6%)	20.8%	0.7%
`*`	142,303 (8.8%)	16.0%	0.8%
`vpn`	32,377 (2.0%)	50.8%	0.2%
`www`	147,588 (9.1%)	4.3%	0.9%
Other	1,131,862 (70.1%)	9.3%	6.7%

Table 2: **Certificates Missing from CT** — Some issuers, such as GoDaddy, have their certificates appear in CT at a lower rate than the general population, and the same is true of `mail.`, `*.`, and `vpn.` subdomain certificates.

	Certificates missing from Censys	Fraction missing	Fraction of universe
All	10,540,648 (100.0%)	62.0%	62.0%
Let's Encrypt	4,401,674 (41.8%)	90.8%	23.5%
CloudFlare	2,381,940 (22.6%)	81.3%	13.9%
Other	3,757,050 (35.6%)	40.7%	22.1%

Table 3: **Certificates Missing from Censys** — Let's Encrypt reports all certificates to CT, even those never served. Cloudflare certificates are only served with SNI.

		FQDNs from CT
With SNI	Accepted connection	20,305,155 (100%)
Without SNI	Accepted connection	15,598,532 (77%)
	Same certificate as SNI	7,021,206 (35%)
	Different certificate	8,577,326 (42%)

Table 4: **SNI Behavior** — 77% of active domains extracted from CT logs accepted connections without SNI, but only 35% served the same certificate as when contacted with SNI.

3.1 Limits of Certificate Transparency

While CT is by far the largest perspective, it still misses 9.5% of the certificate universe, including 29.6% of GoDaddy certificates and 28.4% of cPanel certificates, as shown in Table 2. None of the CAs in the table submit domain validated (DV) certificates to public logs. In contrast, CT captures 99.3% of CloudFlare certificates and 100% of Let's Encrypt certificates.

We also find that CT is skewed towards web content and away from other TLS-based services, such as webmail, that might not be linked to by websites Google crawls. For example, we find that CT misses 20.8% of certificates with

# seen in Alexa Top 1M		... in both CT and IPv4 scans	... in CT scan only	... in IPv4 scan only	... in neither
Certificates	288,220	203,842 (70.7%)	76,085 (26.4%)	7,236 (2.51%)	1,057 (0.37%)
FQDNs	1,906,302	663,129 (34.8%)	1,229,484 (64.5%)	11,769 (0.62%)	1,920 (0.10%)
Sites	916,789	327,894 (35.8%)	583,474 (63.6%)	4,663 (0.51%)	758 (0.08%)

Table 5: **Coverage of Alexa Results from CT Scans and IPv4 Scans**—IPv4 scanning misses 26% of certificates, 65% of FQDNs, and 64% of sites found in our Alexa scans, but combining IPv4 and CT scans yields >99% coverage for each category.

# seen in any perspective		... in both CT logs and Censys	... in CT logs only	... in Censys only	... in neither
Certificates	16,989,236	4,927,174 (29.0%)	10,447,762 (61.5%)	1,521,414 (8.96%)	92,886 (0.55%)
FQDNs	32,454,061	12,156,237 (37.5%)	17,817,430 (54.9%)	2,379,463 (7.33%)	100,931 (0.31%)
Sites	12,673,514	4,412,464 (34.8%)	7,794,002 (61.5%)	426,168 (3.36%)	40,880 (0.32%)

Table 6: **Coverage of Universe in CT Logs and Censys**—Combining the results from CT logs and Censys covers more than 99% of the certificates, FQDNs, and sites that can be found using any of our perspectives.

# seen in our zone scan		... in both CT logs and Censys	... in CT logs only	... in Censys only	... in neither
Certificates	2,431,246	1,283,379 (52.8%)	1,073,965 (44.2%)	37,825 (1.56%)	36,077 (1.48%)
FQDNs	11,881,085	5,231,678 (44.0%)	6,495,535 (54.7%)	65,664 (0.55%)	88,208 (0.74%)
Sites	5,663,431	2,544,950 (44.9%)	3,054,848 (53.9%)	26,409 (0.47%)	37,224 (0.66%)

Table 7: **Coverage of Zone Results from CT and Censys**—CT logs and Censys together cover >98% of our zone scan results.

the subdomain `mail` and 50.8% of certificates with a subdomain containing `vpn`. In contrast, CT only misses 4.3% of certificates with the subdomain `www`.

3.2 Limits of Censys

The Censys snapshot only covers 38% of our certificate universe. Two sources are responsible for approximately 64% of the missing certificates. As shown in Table 3, 90.8% of Let's Encrypt certificates are absent from Censys, accounting for 42% of the certificates missed by Censys and 23.5% of the certificate universe. Let's Encrypt submits all issued certificates to CT, but it appears that many of these certificates are inaccessible without SNI or are not served on public sites.

CloudFlare accounts for 17% of all certificates in this study, but Censys misses 81% of these, resulting in an exclusion of 13.9% of the certificate universe. We manually confirmed that the vast majority of CloudFlare certificates are only accessible through SNI. This intuitively makes sense because most Censys certificates are found through IPv4 scans that do not include SNI information.

3.3 Sites Requiring SNI

In order to directly measure the impact of SNI, we performed two scans over all FQDNs contained in valid certificates in the CT logs. We scanned 30 million domain names, and were able to complete successful HTTPS handshakes with 68% using SNI. As shown in Table 4, only 77% of domains that accepted a connection made with SNI accepted connections without it, and only 35% returned the same certificate as when SNI was used. This further shows that scanning without SNI misses a substantial fraction of websites.

In order to understand if this discrepancy applies to commonly visited sites, we can limit the scope of our comparison to certificates discovered through Alexa Top Million scanning, as shown in Table 5. Even for these popular sites, IP-based scanning misses 27% of certificates and 65% of sites, due to a lack of SNI—a massive difference compared to the 0.7% that Durumeric et al. found in 2013 [11]. Notably, scans of CT and IPv4 combined provide 98.5% of the certificates presented in our Alexa Top Million scan.

3.4 Passive Traffic Monitoring

The ICSI Notary perspective is derived from passive monitoring of network traffic. It only includes certificates actually seen on the wire, and therefore differs significantly from our other perspectives.

In contrast to our active scans, passive monitoring contains certificates from IPv6 connections. We encountered 822,338 server IP addresses in our Notary dataset. Of these, 8.2% (67,725) were IPv6 addresses, comprising 13% of the observed connections. There were 4,512 certificates that were only encountered on IPv6 addresses, but only 218 of them were not observed in any other perspective. This suggests that IPv6 does not impact conclusions drawn from scanning significantly, but as our passive dataset is relatively restricted, further measurement is necessary to verify this claim. Underscoring the importance of SNI discussed in Section 3.3, only 9.7% of connections in the Notary dataset did not use SNI. In total, we saw 3,246,725 unique SNI values.

The Notary saw only 3,805 certificates that were not observed by any other perspective. We believe these are due to certificate changes during the longer passive measurement interval. This is supported by the fact that only 34% of the certificates were encountered at all during the last week of the measurement interval. Furthermore, 75% were issued by CloudFlare, which rotates certificates quickly.

Alexa Top 1M		Common Crawl		CT Scan		IPv4 Scan		Zone File	
Comodo	30%	Let's Encrypt	24%	Let's Encrypt	41%	Comodo	17%	Let's Encrypt	28%
GeoTrust	17%	Comodo	21%	Comodo	15%	GoDaddy	16%	Comodo	24%
GoDaddy	10%	GeoTrust	15%	GeoTrust	9%	GeoTrust	15%	GeoTrust	12%
Let's Encrypt	8%	GoDaddy	9%	GoDaddy	7%	GlobalSign	7%	GoDaddy	12%
GlobalSign	7%	cPanel	7%	GlobalSign	5%	Let's Encrypt	7%	cPanel	7%
Other (351)	29%	Other (475)	25%	Other (560)	23%	Other (567)	62%	Other (377)	18%

Table 8: **Top Certificate Issuers** — The most common issuers differ depending on the perspective studied. Let's Encrypt has its highest popularity in CT Logs, cPanel only appears in the top five for zone file scanning, and IPv4 scanning yields a longer tail.

There were 68,700 certificates seen by our passive measurement that were not present in Censys. They have a similar composition to the certificates in CT logs that Censys missed. 39% are issued by Let's Encrypt, which requires sites to frequently re-issue certificates. 20% are attributable to Wordpress-hosted blogs and 13% to CloudFlare, services that heavily depend on SNI and would therefore not be seen by IPv4 scans.

3.5 Combining CT and Censys

As Table 6 shows, combining data from CT logs and Censys yields 99.4% coverage of all certificates observed by any of our perspectives and 99.7% of all FQDNs and all sites. However, since these perspectives are also our two largest data sources, this statistic may be artificially inflated.

Fortunately, scanning all domains in the `.com`, `.net`, and `.org` zone file provides us with ground truth for all sites being served on the root and `www` subdomain in those zones. We can use this to understand what coverage each perspective gives in these zones. While the zones do not contain all subdomains or even all domain names on the Internet, they are a large subset: 153 million unique domains. We compare Censys data and CT logs over the certificates obtained from the zone scans in Table 7.

Of these certificates, we find 98.5% are obtained through either Censys or CT logs. Since this is smaller than the corresponding percentage for all certificates we observe, there are likely certificates being hosted in other zones and in other subdomains not observed by any method we use. Therefore, the coverage of IPv4 scans and CT logs on the entire population of certificates on the Internet is overestimated by Table 6.

Conversely, the coverage of Censys scanning on the zone dataset is increased to 54%, from 38% over all certificates observed. This is potentially due to certificates in CT logs that are not actively hosted on the Internet (e.g., intranet sites). As a result, the Censys coverage of the entire population of certificates on the Internet is underestimated by Table 6.

4. IMPACT ON HTTPS RESEARCH

A number of recent studies have used IPv4 and Alexa Top Million scanning to measure how HTTPS is deployed in the wild [4, 7, 8, 10, 11, 14, 22, 28, 31, 32]. Our finding that IPv4 scans miss nearly two-thirds of certificates suggests that, if these studies were performed today, they might not accurately reflect the state of the Internet.

	IPv4 Addresses	FQDNs	Sites
IPv4 Scan	1.88%	4.14%	5.18%
Common Crawl	4.54%	1.42%	1.51%
Zones	3.91%	0.90%	0.96%
Alexa	2.71%	0.90%	0.96%
CT Scan	3.73%	1.12%	1.18%

Table 9: **Rate of Vulnerability to FREAK** — Vulnerability rates measured by each methodology vary significantly.

To provide a concrete example of the differences caused by different perspectives, we present a survey of sites vulnerable to the FREAK attack [28] in Table 9. The data comes from scanning each of: the IPv4 address space, CT log FQDNs, Alexa Top Million domains, our zone files, and domains extracted from Common Crawl. For each set, we measure how many responsive hosts are vulnerable.

The number of vulnerable hosts changes with each perspective we measure. The vulnerability rates range from 1.88% of IPs vulnerable when measured by IPv4 scanning, to 4.54% of IPs vulnerable when measured by our Common Crawl scan. This variation demonstrates the necessity of considering the perspective used when performing measurements.

We also observe differences when comparing the most common certificate issuers seen in each perspective, as shown in Table 8. We measure this by grouping certificates by their issuer organization field and manually deduplicating similar issuer names. Different perspectives display differing views on which CA is most popular. For example, Let's Encrypt is the most popular CA in CT, Common Crawl, and zone scans, but it is ranked fifth in IPv4 scans.

5. RELATED WORK

This work was inspired by a large body of research focusing on the HTTPS ecosystem and supporting PKI [4, 7, 8, 10, 11, 14, 28, 31, 32]. These works have run the gamut of HTTPS measurement, ranging from the certificate authority ecosystem [11, 14] to cryptographic keys generated without entropy [16], and how operators react to vulnerabilities [10].

In 2010, the EFF launched the SSL Observatory [14], in which they performed a scan of the IPv4 address space over a three month period in order to identify trusted certificate authorities. Later, in 2011, Vratonjic et al. [31] crawled the Alexa Top Million finding that only 5.7% of websites correctly deploy HTTPS. The same year, Holz et al. [18]

carried out a similar study, combining active measurements of the Top Million from from several vantage points, with data from passive measurement at a large research institution. They briefly compared the differences of their vantage points, concentrating on differences caused by scan origins.

In 2013, Durumeric et al. [11] analyzed the state of the HTTPS PKI by repeatedly scanning the IPv4 address space. They briefly considered how much of the Alexa Top Million was only accessible using SNI, finding that, at that time, SNI was not widely required. Other studies use data based on websites on the Top Million [19], and scans of the IPv4 ecosystem [10, 15, 16]. Studies have also been based on combinations of Alexa sites, random domains, known phishing domains data from passive network monitoring [1, 2, 10, 24, 29] Most recently, studies have also investigated TLS deployment outside of HTTPS [9, 17].

Our work does not focus on the questions asked by these individual studies, but instead focuses on *how* these types of studies should be measuring the HTTPS ecosystem. We hope that by better validating different methodologies for studying the HTTPS PKI, we can help future papers obtain more accurate measurements.

6. CONCLUSION

Over the past five years, dozens of studies have measured the HTTPS ecosystem and supporting PKI. Unfortunately, without a clear ground truth, these studies pieced together a view built on a series of fractured and imperfect methodologies. In this work, we investigated these methodologies, finding that IPv4 enumeration no longer provides a representative view of how TLS servers are configured, due to SNI deployment. IPv4 scans miss more than two-thirds of valid certificates and associated measurements can differ dramatically from site-based approaches. Certificate Transparency provides a new perspective, which finds 90.5% of certificates observed in this study, but is skewed towards a few authorities that submit the certificates they issue. We find that a more comprehensive yet readily accessible methodology is to use a combination of CT and Censys data, which together account for 99.4% of our observed certificates. To this end, we are working with the Censys team to implement continuous certificate synchronization between Censys and Google's CT logs, which will soon make either data source a nearly comprehensive view of trusted HTTPS certificates.

Acknowledgements

The authors thank David Adrian, Ryan Hurst, and Ben Laurie for insightful discussions and feedback. We thank the exceptional sysadmins at the University of Michigan for their ongoing help and assistance, and Charlie Mattison for repeatedly going above and beyond in support of our research. We also thank the anonymous reviewers and our shepherd, Dave Levin. This material is based upon work supported by the National Science Foundation under grants CNS-1345254, CNS-1409505, CNS-1505790, CNS-1518741, CNS-1518888, CNS-1530915, CNS-1528156, and ACI-1348077, by the Google Ph.D. Fellowship in Computer Security, and by an Alfred P. Sloan Foundation Research Fellowship.

7. REFERENCES

[1] D. Akhawe, J. Amann, M. Vallentin, and R. Sommer. Here's my cert, so trust me, maybe? Understanding TLS errors on the web. In *22nd International World Wide Web Conference*, May 2013.

[2] J. Amann, R. Sommer, M. Vallentin, and S. Hall. No attack necessary: The surprising dynamics of SSL trust relationships. In *29th Annual Computer Security Applications Conference*, Dec. 2013.

[3] J. Amann, M. Vallentin, S. Hall, and R. Sommer. Extracting certificates from live traffic: A near real-time SSL notary service. Technical Report TR-12-014, ICSI, Nov. 2012.

[4] B. Beurdouche, K. Bhargavan, A. Delignat-Lavaud, C. Fournet, M. Kohlweiss, A. Pironti, P.-Y. Strub, and J. K. Zinzindohoue. A messy state of the union: Taming the composite state machines of TLS. In *36th IEEE Symposium on Security and Privacy*, May 2015.

[5] Certificate Transparency: Extended validation in Chrome. https://www.certificate-transparency.org/ev-ct-plan.

[6] Common Crawl. https://commoncrawl.org/.

[7] The DROWN attack. https://drownattack.com/.

[8] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman. Censys: A search engine backed by Internet-wide scanning. In *22nd ACM Conference on Computer and Communications Security*, Oct. 2015.

[9] Z. Durumeric, D. Adrian, A. Mirian, J. Kasten, E. Bursztein, N. Lidzborski, K. Thomas, V. Eranti, M. Bailey, and J. A. Halderman. Neither snow nor rain nor MITM... an empirical analysis of email delivery security. In *15th ACM Internet Measurement Conference*, Oct. 2015.

[10] Z. Durumeric, J. Kasten, D. Adrian, J. A. Halderman, M. Bailey, F. Li, N. Weaver, J. Amann, J. Beekman, M. Payer, and V. Paxson. The matter of Heartbleed. In *14th ACM Internet Measurement Conference*, Nov. 2014.

[11] Z. Durumeric, J. Kasten, M. Bailey, and J. A. Halderman. Analysis of the HTTPS certificate ecosystem. In *13th ACM Internet Measurement Conference*, Oct. 2013.

[12] Z. Durumeric, E. Wustrow, and J. A. Halderman. ZMap: Fast Internet-wide scanning and its security applications. In *22nd USENIX Security Symposium*, Aug. 2013.

[13] D. Eastlake. Transport Layer Security (TLS) Extensions: Extension Definitions. RFC 6066, 2011.

[14] Electronic Frontier Foundation. The EFF SSL observatory. https://www.eff.org/observatory.

[15] X. Gu and X. Gu. On the detection of fake certificates via attribute correlation. In *Entropy*, Nov. 2014.

[16] N. Heninger, Z. Durumeric, E. Wustrow, and J. A. Halderman. Mining your Ps and Qs: Detection of

widespread weak keys in network devices. In *21st USENIX Security Symposium*, Aug. 2012.

[17] R. Holz, J. Amann, O. Mehani, M. Wachs, and M. A. Kaafar. TLS in the wild: An Internet-wide analysis of TLS-based protocols for electronic communication. In S. Capkun, editor, *Network and Distributed System Security Symposium*, Feb. 2016.

[18] R. Holz, L. Braun, N. Kammenhuber, and G. Carle. The SSL landscape: A thorough analysis of the x.509 PKI using active and passive measurements. In *11th ACM Internet Measurement Conference*, Nov. 2011.

[19] L.-S. Huang, S. Adhikarla, D. Boneh, and C. Jackson. An experimental study of TLS forward secrecy deployments. *IEEE Internet Computing*, 18(6), 2014.

[20] J. C. Jones. 124 days of Let's Encrypt, Apr. 2016. https://tacticalsecret.com/124-days-of-lets-encrypt/.

[21] B. Laurie, A. Langley, and E. Kasper. Certificate Transparency, 2013. http://www.certificate-transparency.org/.

[22] A. K. Lenstra, J. P. Hughes, M. Augier, J. W. Bos, T. Kleinjung, and C. Wachter. Public keys. In *32nd International Cryptology Conference*, Aug. 2012.

[23] Let's Encrypt: Certificates. https://letsencrypt.org/certificates/.

[24] M. A. Mishari, E. D. Cristofaro, K. M. E. Defrawy, and G. Tsudik. Harvesting SSL certificate data to mitigate web-fraud. *CoRR*, abs/0909.3688, 2009.

[25] Mozilla Foundation. Public suffix list. https://publicsuffix.org/.

[26] R. Sleevi. Sustaining digital certificate security, Oct. 2015. https://security.googleblog.com/2015/10/sustaining-digital-certificate-security.html.

[27] StartCom log all issued SSL certificates to public CT log servers, Mar. 2016. https://www.startssl.com/NewsDetails?date=20160323.

[28] Tracking the FREAK attack. http://freakattack.com/.

[29] N. Vallina-Rodriguez, J. Amann, C. Kreibich, N. Weaver, and V. Paxson. A tangled mass: The Android root certificate stores. In *10th ACM Conference on Emerging Networking Experiments and Technologies*, 2014.

[30] VeriSign: Zone file information. https://www.verisign.com/en_US/channel-resources/domain-registry-products/zone-file/index.xhtml.

[31] N. Vratonjic, J. Freudiger, V. Bindschaedler, and J.-P. Hubaux. The inconvenient truth about web certificates. In *10th Workshop on Economics in Information Security*, 2011.

[32] L. Zhang, D. Choffnes, D. Levin, T. Dumitras, A. Mislove, A. Schulman, and C. Wilson. Analysis of SSL certificate reissues and revocations in the wake of Heartbleed. In *14th ACM Internet Measurement Conference*, Nov. 2014.