Accept the Risk and Continue: Measuring the Long Tail of Government https Adoption

Sudheesh Singanamalla
University of Washington
sudheesh@cs.washington.edu

Esther Han Beol Jang
University of Washington
infrared@cs.washington.edu

Richard Anderson
University of Washington
anderson@cs.washington.edu

Tadayoshi Kohno
University of Washington
yoshi@cs.washington.edu

Kurtis Heimerl
University of Washington
kheimerl@cs.washington.edu

ABSTRACT

Across the world, government websites are expected to be reliable sources of information, regardless of their view count. Interactions with these websites often contain sensitive information, such as identity, medical, or legal data, whose integrity must be protected for citizens to remain safe. To better understand the government website ecosystem, we measure the adoption of `https` including the "long tail" of government websites around the world, which are typically not captured in the top-million datasets used for such studies. We identify and measure major categories and frequencies of `https` adoption errors, including misconfiguration of certificates via expiration, reuse of keys and serial numbers between unrelated government departments, use of insecure cryptographic protocols and keys, and untrustworthy root Certificate Authorities (CAs). Finally, we observe an overall lower `https` rate and a steeper dropoff with descending popularity among government sites compared to the commercial websites & provide recommendations to improve the usage of `https` in governments worldwide.

CCS CONCEPTS

• **Networks** → **Network measurement**; **Network security**; *Public Internet*; • **Applied computing** → *Computing in government*; • **Social and professional topics** → Governmental regulations.

KEYWORDS

TLS, HTTPS, Measurement, Government, Certificate Authorities, X.509 Certificates

ACM Reference Format:
Sudheesh Singanamalla, Esther Han Beol Jang, Richard Anderson, Tadayoshi Kohno, and Kurtis Heimerl. 2020. Accept the Risk and Continue: Measuring the Long Tail of Government `https` Adoption. In *ACM Internet Measurement Conference (IMC '20), October 27–29, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3419394.3423645

1 INTRODUCTION

Today, most secure web communication takes place over HyperText Transfer Protocol Secure (`https`). Using Transport Layer Security (TLS) to encrypt `http` requests and responses, `https` provides users with message *authentication*, *integrity* and *confidentiality*. Many elements of `https` usage have been explored, with previous work focused on measuring the cost of `https` [59], analyzing the certificate ecosystem [25], and more recently, examination of `https` adoption in the web of 2017 [31].

While the most recent measurements by Felt et al. [31] focused on adoption of `https` using the Alexa top 1 Million dataset, many critical web resources are unlikely to fall within this dataset, such as websites run by local and national governments. Such sites, often serving smaller geographic regions or countries without a large web presence, are trusted with holding sensitive user data for civic functions or providing information such as local infectious disease numbers. Prior case studies have shown that citizens visit local county government websites for a wide range of services including job openings, local demographics, budgets, meeting minutes, details of contracts and their summaries, and for official contact information of their elected representatives [10]. Research also shows that websites providing quality e-services help build trusted relationships between citizens and their governments; further, low-traffic local government websites such as utilities, water *etc.*, while not present in top million lists, are still actively used in citizens' daily lives [76]. Attackers therefore may target government sites to disrupt critical infrastructure, steal identifying data, disenfranchise citizens and influence politics, or decrease their trust in the government. Providing secure access to local ".gov" sites should be of high priority for governments.

Despite the high importance of government websites' integrity, we find that greater than 70% of the total government websites measured worldwide (in a scan of 135,408 of which only 12,293 (9.07%) are in standard top millions lists), do not use valid `https`. Of the 53,256 (39.33%) websites that attempt to support `https`, 15,223 (28.58%) are invalid with a large variety of certificate errors. We identify major categories and frequencies of these errors, including ∼5.50% expired certificates, ∼13% use of insecure cryptographic protocols, ∼15% use of self signed certificates either at the leaf level or in the cert chain, and 1,390 instances of public key reuse between unrelated governments. Including the websites that implement `https` correctly but do not enforce upgrades to `https`, this number rises to 19,349 (36.33%).

IMC '20, October 27–29, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8138-3/20/10…$15.00
https://doi.org/10.1145/3419394.3423645

We make the following contributions: (1) perform a large-scale study of https adoption of global government websites including outside of the "top millions"; (2) identify trends in type of hosting, cryptographic key usage, CAs, and corresponding https validity; (3) perform in-depth case studies on two countries, the United States of America (USA) and South Korea (ROK), selected for the existence of *authoritative* government hostname lists, and compare them; (4) responsibly disclose these weaknesses, and measure our influence on https validity; (5) discuss the limitations of our measurements, and (6) conclude with recommendations to stakeholders to encourage & improve https adoption.

2 RELATED WORK

2.1 Datasets & Tools

Datasets of websites exist on the Internet for research use, including the Alexa million datasets which rank sites by popularity, the Cisco million [16] which ranks by traffic volume, and the Majestic million [42]–an open source version of the Alexa million since its acquisition by Amazon. Tranco, another public list, attempts to provide a more stable ranking for web measurement avoiding the flux of prior datasets [49]. Le Pochat *et. al.* note that only 49% of the domains in the Umbrella datasets are available, responding with a success status code of 200, as are only 89% of the Majestic million [49]. Our work uses these datasets as a seed set, which we then expand through web crawling, Amazon Mechanical Turk tasks, and hand-searching domains to increase the number of unique measurable government websites from 27,532 to 135,408. This is a substantial increase from government websites in existing datasets, and forms the basis for our analysis. In prior work, tools like ZMap and CFSSL have enabled researchers to perform large scale studies on Internet hosts [19, 22]. Services like Censys constantly monitor the Internet for https adoption in addition to detecting vulnerabilities like heartbleed [23].

2.2 https Measurements

The closest and most recent work to our analysis is by Felt *et al.* on measuring adoption of https across the web [31]. Google, in their report indicated a modification to their search algorithm to boost sites using https [11]. Our analysis is different in that prior efforts largely focus on the "head" of the Internet, *i.e.* popular domains as found in top million lists. However, in this paper, we explicitly include the "long tail" of government websites as they are especially critical to users' safety but do not commonly appear in the top million lists. Mirian *et al.* similarly measured https among general sites outside the top millions, finding that services providing free certificates such as *Let's Encrypt* improve overall adoption of https and that general web domains also use *Let's Encrypt* four times more than other CA authorities [55]. We show that Let's Encrypt is also the most popular CA used by government sites globally, though not in every country.

Prior studies have tried to understand the root causes of https certificate errors in Chrome [5] and analyze trust models in CAs [6, 29]. Others focus on challenges in the certificate ecosystem, the need to make them more auditable, and ways that CAs could be incentivized using insurance models with benefits negotiated between CAs and domains [24, 25, 38, 52, 64].

Certificate Transparency (CT) is one such effort to make issuance more auditable by continuously recording SSL/TLS certificates on an append-only database [48, 70]. Previous studies used CT logs to detect phishing domains which were issued certificates, and phishing attackers using a honeypot [70]. Another study in 2016 measured worldwide adoption of Let's Encrypt-issued certificates, which are automatically published to CT logs [80], and noticed that adoption was higher in countries with high Internet penetration [51]. While CT provides the largest view of certificates on the Internet, it misses around 10% in the .com, .net and, .org zones [80]. There is no existing measurement of the number of government domain certificates missing from CT logs.

2.3 Usable Security

Multiple studies have shown users' proclivity to ignore warning messages displayed by browsers when interacting with websites on the Internet [47, 73]. Studies of webmasters have shown that they often unknowingly misconfigure SSL/TLS certificates on web servers, but are split on the importance of https and sometimes even misconfigure certificates on purpose [30]. Many end-users misinterpret https on websites as indicators of a secure site [68]. Understanding TLS errors, communicating the dangers of non-https websites, and understanding challenges in https adoption have been studied [6, 17, 30, 46]. Given our result that government websites have different (and worse) https adoption properties than commercial websites, studies focused on government webmasters could be an interesting future direction.

3 HTTPS BACKGROUND

3.1 HTTPS, Certificates and the Web

https is an encrypted data transfer protocol between a web browser client and a web server providing a secure version of the older http protocol. https uses Transport Layer Security (TLS), a successor to the now deprecated Secure Sockets Layer (SSL) protocol, to establish secure communication using asymmetric key cryptography. A TLS handshake starts the process of establishing a secure connection to a website. The handshake begins with the client and server negotiating the TLS version and cipher suite to use. The client validates the certificate provided by the server, then generates a *premaster secret* which is encrypted with the server's public key. This is used to establish a shared session key, enabling an encrypted communication channel [18].

TLS/SSL certificates are specific files hosted by web servers containing the domain hosts' public key along with identity information, such as the domain name they wish to use and the name of the organization. These certificates are (per best practice) cryptographically signed (attested) and issued for a fixed duration by a trusted certificate authority (CAs). CAs previously voted to limit certificate lifetimes to 825 days [13], and recently further reduced lifetimes to one year starting September 2020 [14, 33]. However, it is also possible to create a *self-signed certificate* which is cryptographically valid but not attested by a CA, thus limiting the privacy benefits.

If configured correctly, https allows communication to remain confidential and non-tamperable, providing an authenticated medium between client and server with the assurance that communications are only being received and read by the intended recipient. Popular projects like Let's Encrypt, a US non-profit, have made it possible for website operators to add https support for absolutely no fee [4]. Cloud providers like Azure, Google Cloud, along with Content Delivery Network (CDN) providers like Cloudflare and Akamai, have similarly made it easy to integrate https by intercepting and handling the requests [26] in a secure, easy to configure manner.

http, without https, enables *man-in-the-middle* (MITM) attacks wherein the adversary can eavesdrop, secretly alter, and relay communications between two parties, sending its communication entirely in plaintext. Expert attackers, malicious governments, or Internet Service Providers could proxy requests and show modified content to the user, steal their information, or use it for surveillance. Such attacks have been thoroughly studied and publicly documented by cybersecurity companies [2] and organizations such as the Open Web Application Security Project (OWASP) [20]. The lack of a matching root certificate during validation of a certificate chain results in an error indicating undetected local issuer certificate [63].

3.2 Certificate Authorities

Certificate Authorities (CAs) are trusted third parties whose core responsibility is to issue SSL/TLS certificates. CAs and their certificates are treated as trust anchors and shipped by default by software providers (usually with browsers or operating systems) such as Microsoft, Google, Apple, and Mozilla [9, 54, 57]. The list of default trusted root CAs can differ between browsers and tools. Our analysis of the trust stores show that Apple includes 174 default root trusted certificates, while Microsoft [54] includes 402 default root certificates. The Mozilla NSS [57] trusted certificate store consists of 152 default root trusted certificates. NSS trusts 52 individual root CA owners, while Microsoft and Apple trust 133 and 69 root CA owners respectively. Any valid intermediate CA must be authorized as a CA. Therefore, a weak CA in a certificate's chain of trust is a weak link in a website's security, exemplified by the compromises of DigiNotar and Comodo [7, 8].

A certificate issued by a CA binds the public key of the web host to the domain name and is cryptographically established by the CA signing the contents with its private key. A CA responds to a request to issue a certificate by challenging the domain host to prove its ownership. Such Domain Validated (DV) certificates are the most common type. CA-issued certificates can also include information such as organization names, postal address, or an administrator email address. These Extended Validated (EV) certificates are rigorously validated by the CA before issuance and are intended to make phishing attacks with valid certificates harder.

EV certificates (limited to a 2 year validity [34]) were previously treated in a special manner by most browsers, *e.g.* by displaying the name of the business entity along with the green lock symbol indicating availability of https [74]. However, it was still possible for a malicious attacker to register a company with the same name in a different physical address and request an EV certificate. EV certificates are generally expensive, with a fee for issuance. They have been widely adopted by large Internet companies, payment gateways, and banks providing online services. However, their popularity has reduced due to concerns about their effectiveness and the move by major browsers to avoid distinguishing visually between EV and DV certificates in the interface [74].

4 METHODOLOGY

4.1 Seed Dataset

Throughout the work, we use the term "hostname" to refer to the full subdomain+domain strings identifying unique websites (*e.g.* "blog.example.com" with subdomain "blog" and domain "example.com"), rather than URL strings which may include subdirectories. We also define website or hostname "availability" as successful resolution of the DNS query and a 200 code in response to a web request to load page content.

We begin by generating an initial "seed" list of government hostnames by merging the publicly available top-million datasets mentioned in section 2.1, including the Majestic Million dataset, Cisco top 1 Million dataset, one historical copy of the Alexa top 1 Million dataset published in August 2019, and the Censys research dataset produced by the University of Michigan and made available through Google BigQuery [23, 69]. This merged dataset of hostnames is then filtered and de-duplicated to include only government websites through the method described in 4.1.1.

As of August 2019, this yielded a seed list of 27,532 unique government hostnames. An initial query using the Majestic Million dataset indicated that nih.gov is the highest-ranked government hostname (51st), and ncb.gov.sg is the lowest (999,825th). The top government website that does not have an TLS/SSL certificate, ranked at 222, is miit.gov.cn and belongs to the Ministry for Industry and IT of the People's Republic of China.

4.1.1 Government Hostname Filtering. We separate government and non-government sites through a regular expression filter for hostnames using standard government formats. A popular format used by many countries is *.gov.country-code*, and all countries except the United States use only one domain extension. However, the USA uses both *.gov.us* and *.gov* for official government purposes, in addition to a dedicated federal *.fed/.fed.us* and military *.mil* top level domain (TLD) without the "us" country code.

Government domain names and extensions depend heavily on countries' primary languages. Countries with French as a primary language often use *.gouv*, and those with Spanish use *.gob* followed by country code. Kenya, Indonesia, Japan, Korea, Thailand and Uganda use *.go* followed by the country code. Some countries use *.gub, .govern, .government*, and *.guv*, New Zealand uses *.govt* and Switzerland uses *.admin*. We filter hostnames in the dataset using these known expectations and exceptions, along with country code extensions, as a conservative filter with high precision but limited recall. This was decided to ensure that our list was comprised of only government websites. For example, *environment.gov.au*, *geoportal.capmas.gov.eg*, *stats.data.gouv.fr* & *www.pwebapps.ezv.admin.ch* are valid hostnames because they follow the format of a valid government domain name extension followed by a country code, making them valid ccTLDs included in our scan.

4.2 Expanding the Dataset

We expanded this initial list through three separate mechanisms: 1) crowdsourcing local hostnames using Amazon Mechanical Turk, 2) crawling the hostnames in our list, and 3) hand-curating and whitelisting a set of government hostnames which do not use standard government domain extensions.

4.2.1 Amazon Mechanical Turk (MTurk). Seeding with sites from the top millions inherently biases our results towards larger or more connected countries. To combat (but not entirely remove) this bias, we used Amazon's Mechanical Turk (MTurk), a popular crowdwork platform [44], to publish tasks for finding government websites for countries where we had only a few or no hostnames. Each task asked a worker to enter up to six URLs from a specific country, with USD 0.60 paid per task. To encourage site diversity, we asked workers to find different categories of government sites. The categories were: the National Government (or the Presidency if no national government site was available), Public Health (or a government News/Media site if none available), Taxes (or Finance Ministry if none available), Immigration or Travel, and any 2 different departments not covered. The tasks were completely anonymous with no repeat responses allowed from the same worker. The only demographic information queried was a binary Yes/No indicating if the worker was from the country in the issued task.

We published tasks for countries with less than 11 hostnames in the seed list (from section 4.1), including Andorra, Chad, Chile, Democratic Republic of the Congo (DRC), Costa Rica, El Salvador, Guatemala, Iceland, New Zealand, Nicaragua, Panama, Tanzania, Thailand, Tonga, Greenland, Western Sahara, Falkland Islands, Puerto Rico, New Caledonia, Solomon Islands, Northern Cyprus, Somaliland, Kosovo, South Sudan, and Niger. We received 108 responses, of which we accepted 75 after manual inspection. 11 workers self-reported as being from one of these countries. They were: 4 from Greenland, 2 from the Democratic Republic of the Congo (DRC), and 1 each from Andorra, Costa Rica, New Zealand, New Caledonia, Solomon Islands and Kosovo.

We obtained a total of 199 unique hostnames from the 108 MTurk tasks we issued, with 61 already in the seed list. 138 new hostnames were added to our seed list, bringing the size to 27,794.

4.2.2 Crawling Government Websites. We built a web crawler for the above seed list (inclusive of added MTurk hostnames) that visits every hostname, gathers all links on the page not yet seen by the crawler with a valid country code extension (according to ICANN [40]) and follows the links for 7 levels of depth before terminating the crawl for that hostname.

The crawler began with 27,794 hostnames and retrieved 843,561 hostnames in total, resulting in 301,219 unique hostnames after deduplication, of which only 7,723 were repeated from the top million datasets. 134,812 remained after strict filtering for government hostnames as described in Section 4.1.1. The crawls were completed from the University of Washington between 1st-3rd March 2020.

We measured the rate at which the dataset grew from our initial seed list as a result of the crawler. The rate of hostname discovery steadily declines for each level after the 5th, leaving us with 134,812 unique government hostnames at the end of the crawl. See the appendix A.3 for more information on the growth of the dataset.

4.2.3 Hostname Search and Whitelisting. Finally, we manually investigated the seed list for each and every country, adding missing websites to ensure inclusion of improperly filtered hostnames, obvious sites from top search engine results, and long-tail countries still having less than 11 total sites after the MTurk tasks. We found these websites via a combination of Google search, manual crawling of seed list links and foreign embassy or non-government travel sites, and careful individual scrutiny for signs of legitimacy as well as impersonation or phishing (to the best of the authors' ability and expertise). This produced a hand-curated whitelist of 596 government hostnames from 62 countries, which we included with the final list of 134,812 filtered unique hostnames, resulting in a total of 135,408. Even after this process, 15 countries remained with less than 11 sites: Chad, Comoros, DRC, Equitorial Guinea, Eritrea, Honduras, Nauru, Niger, North Korea, Palau, Sao Tome and Principe, South Sudan, Togo, and Tuvalu.

We also manually added hostnames from Germany, Greenland, Gabon, Denmark, and the Netherlands, which do not use any variation of our expected government domain extensions, as well 14 countries using TLDs such as .com, .org, and .net, to our whitelist. We did not crawl these whitelisted hosts with our automated crawler because we could not programmatically confirm linked sites as government-operated without manually visiting and tagging the crawl results.

Using the final list of hostnames, we performed measurements between April 22nd and April 26th, 2020. For the measurements, we performed full TLS and TCP handshakes with the root page of each website and retrieved the certificate chain along with the peer certificate. In case of failures to connect, we performed 3 retries for the hostname by adding the request to the queue. If the host did not return a status 200 code after three attempts, either because the domain name could not be resolved or we could not fetch any content over `http` or `https`, we deemed the website "unavailable" and excluded it from further analysis. The results in this paper were obtained from a single snapshot. Future work could monitor sites periodically to identify changes in `https` adoption. [1]

Number of Govt. Websites	Majestic Million	Cisco Million	Tranco Million
Top 1000 (1K)	56	0	30
Top 10000 (10K)	508	14	373
Top 100000 (100K)	2538	433	2351
Top 1000000 (1M)	12445	9296	12293

Table 1: Overlap of Our Government Website Dataset With Public Top Millions

4.2.4 Ranking. As our authoritative ranking dataset we used the Tranco Million [49], a curated list of top million sites optimized for lower churn and thus more research validity. 12,293 (<10%) of our 135,408 discovered hostnames were present. The small overlap of our generated list and the Tranco million suggest that most of our discovered hostnames likely lie in the long tail of the Internet and outside prior analyses. The overlap with Tranco and other popular top million datasets are presented in Table 1. In Section 5.5 we present comparisons between government and non-government websites in the top million using the Tranco million dataset.

[1] We identified some inaccuracies due to timeouts from our scanners while measuring the adoption of https for New Zealand, Republic of Congo, Togo, and United Arab Emirates. We performed an additional scan on 9/9/2020 and updated our results.

4.3 Certificate Validation

We used *OpenSSL* for validation of certificates and certificate chains downloaded from all of the hosts [62]. To mark a website as valid in our scans, we validate the entire certificate chain. We chose OpenSSL with the default trust store shipped with the Apple Mac operating system [9] imported into the machine over Mozilla's NSS or the Chromium trusted certificate store, since it is the most restrictive and does not include certificates that might be available individually in the browsers' codebases based on their trust with the CA as described in section 3.2. As a result, our scan shows a small number of certificates as invalid which are valid when using a specific browser or operating system, due to our conservative trust store. Based on our disclosure reports and the responses obtained as described in section 7.2.1, we identified 8 hostnames that were invalid in our scans but are valid on some known web browsers and operating systems.

4.4 Ethical Statement

This study was approved by the Institutional Review Board (IRB) and exempted under ID `STUDY00009482` by the University of Washington Human Subjects Division. The authors involved in the study did not tamper with any vulnerable government website and executed a full responsible disclosure process by informing the respective country's government authorities and the corresponding technology or administrative contact listed on the *whois* services of the host. The authors only used port 80 and 443 to access the websites and did not perform any port scanning actions that might result in abuse of the hosts in the target.

5 RESULTS

We provide a number of measurements of `https` adoption among our curated list of government websites. We first present our overall measurements on worldwide adoption of `https` (section 5.1), identify the most popular certificate issuers for government sites (section 5.2), and provide a breakdown of reasons for certificate invalidity among affected hosts (section 5.3). We then identify the effects of hosting type and hosting providers on certificate validity (section 5.4), compare our government websites to non-government websites within the top million (section 5.5), and conclude that government websites have overall poorer valid `https` adoption. To offset data collection biases, we perform detailed case studies with two countries' with official authoritative datasets (section 6) and find certificate invalidity results worldwide.

5.1 `https` Adoption, Use, and Issues

Of 135,408 worldwide government hostnames analyzed, 82,152 (**60.67%**) only support `http`, while 53,256 (**39.33%**) serve their content with `https`. Only 38,033 (**28.08%**) use `https` correctly, even when optimistically including the 4,126 sites that load content on both `http` and `https`.

We show overall results by country as a chloropleth map in Figure 1. Within the United States, while a majority of the websites do support `https`, there are still 1,841 sites (18.45%) that have no `https` and 1,147 sites (11.49%) serving both `http` and `https` traffic; we examine the USA further as a detailed case study in section 6.1.

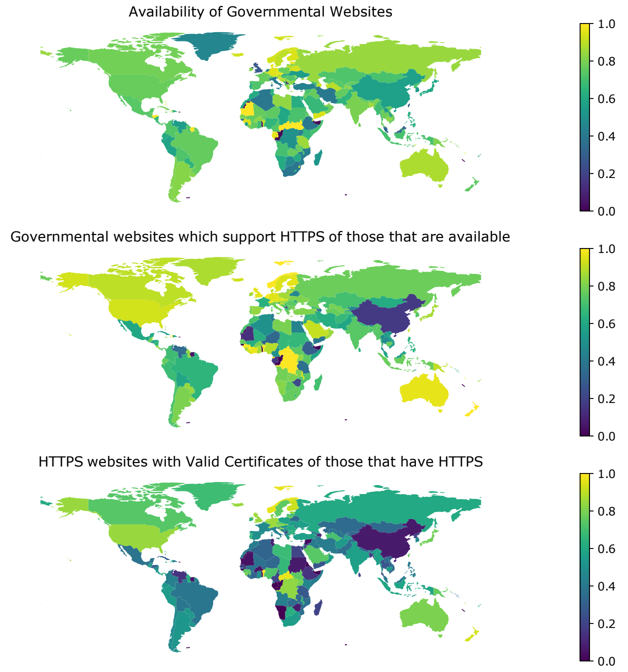
Availability of Governmental Websites
1.0
0.8
0.6
0.4
0.2
0.0
Governmental websites which support HTTPS of those that are available
1.0
0.8
0.6
0.4
0.2
0.0
HTTPS websites with Valid Certificates of those that have HTTPS
1.0
0.8
0.6
0.4
0.2
0.0

Figure 1: Worldwide view of Government Websites

Top: the percentage of government websites from our total list that are available, where the host returns a 200 status code. Middle: the percentage of available sites which support `https`. Bottom: the percentage of sites that support `https` which have valid certificates.

5.2 Certificate Authorities

Most (20.03%) of `https` enabled government websites worldwide use certificates issued by Let's Encrypt with ≈80% of them being valid. ≈20% invalidity is due to expiry, misconfiguration leading to incorrect certificate usage by the host, or self-signing of certificates.

The top 15 CAs used by governments, including Let's Encrypt, do not provide EV certificates. The first major EV certificate issuer, DigiCert, has ∼20% invalid certificates for government hostnames, similar to Let's Encrypt. This case suggests EV certificates obtained for a fee may be equally likely to be invalid as *free* CAs. We show a breakdown of the certificate issuers and their number of invalid certificates worldwide in Figure 2.

The top CAs issuing certificates for government hostnames differ by country. For example, the leading certificate issuer in Switzerland is QuoVadis Global SSL ICA G3, while in China it is Encryption Everywhere DV TLS CA-G1. From a global perspective, Let's Encrypt continues to be the leading CA authority issuing certificates. We expect that this is due to the low cost (free) of certificate issuance and ease of installation with tools like certbot by the Electronic Frontier Foundation (EFF) [27].

5.3 Common Certificate Errors

Combining valid and invalid certificates, 53,256 websites in our list attempt to serve `https` web content. Filtering out 2,721 hostnames which have exceptions and other errors, and 92 hostnames without

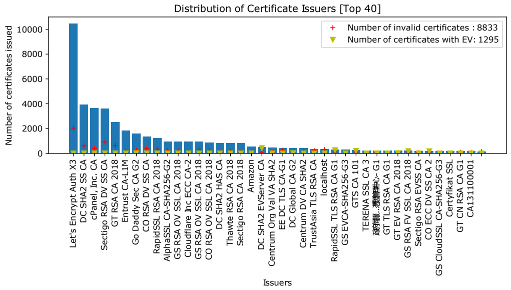

Figure 2: Top 40 Cert Issuers for Government Websites
Abbr: COMODO=CO, DigiCert=DC, GlobalSign=GS, GlobalTrust=GT, Encryption Everywhere=EE, “High Assurance Server”=HAS, “Secure Server”=SS

certificate issuer information encoded in their certificate, we analyze the remaining 50,443 hostnames. 19,781 (**39.21%**) of the sites use a wildcard certificate and 4,486 (**22.67%**) of these are invalid. We further use the EV policy OIDs in Mozilla’s certverifier to check for policy strings corresponding to trusted EV certificates [58], and find 2,145 (**4.24%**) EV certificate hostnames.

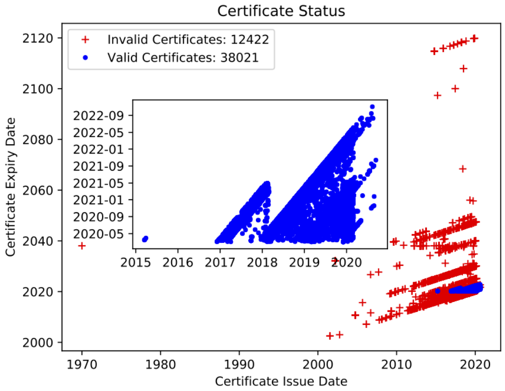

Figure 3: Certificates by issue and expiry date.

The leading cause for certificate invalidity is **host name mismatch**, contributing to **36.6%** of the invalid `https` certificates. Further analysis of some of these mismatches follows in Section 5.3.3. Errors in **retrieving local issuer certificate** and **certificate self-signing** are the next most common. There are instances of government hostnames both using expired certificates and having self-signed certificates in the certificate chain, but this is less than 1% of the hostnames considered. During our scans, 12.7% of the hosts try to negotiate an unsupported SSL protocol (older than SSLv3.0), indicating that the server might be running old unpatched software potentially vulnerable to POODLE [56].

	Count	**%**
Total websites considered	**135,408**	**100**
➤ Content served on HTTP only	82,152	60.67
➤ Content served on HTTPS	53,256	39.33
➤ Valid HTTPS Certificates	38,033	71.41
➤ Invalid HTTPS Certificates	15,223	28.58
➤ Hostname Mismatch	5,571	36.59
➤ Unable to get local issuer cert	3,732	24.51
➤ Exceptions	2,619	17.20
➤ Unsupported SSL Protocol	1,929	73.65
➤ Timed out	378	14.43
➤ Connection refused	135	5.15
➤ Connection Reset by peer	141	5.38
➤ Wrong SSL Version Number	11	0.42
➤ TLSv1 Alert Internal Error	9	0.34
➤ SSLv3 Alert Handshake Failure	7	0.26
➤ TLSv1 Alert Internal Proto. V.	8	0.30
➤ Self-signed certificate	2014	13.22
➤ Certificate Expired	838	5.50
➤ Self-signed certificate in chain	347	2.27
➤ Others	102	0.67

Table 2: Worldwide govt. sites by `https` validity and error
All percentages are computed out of the category level directly above it (for example, Unsupported SSL Protocol accounts for 73.65% of Exceptions.)

5.3.1 Certificate Issue Duration. Valid certificates were commonly issued for a fixed duration of 2-3 years as agreed upon by the CAs [13, 14]. Invalid certificates have a much wider spread in duration (see Figure 3). We find 12,422 total invalid certificates due to hostname mismatches, inability to get local issuer certificates, leaf self-signed certificates, and those in the certificate chain along with expired certificates (excluding those causing exceptions). Only 32% of these had a total validity of less than 2 years. 1,746 (14%) were issued for greater than 3 years. 40 certificates had an expiry date 100 years from the year of issue. 617 websites had invalid certificates issued for 10 years, 155 for 20 years, 36 for 30 years, and 1 for 50 years. 1 certificate had an issue date in 1970 (Unix epoch time) expiring in 70 years, likely indicating misconfiguration. 5,372 (43.24%) were issued for a duration in multiples of 365.

5.3.2 Cryptographic Key Usage & Signing Algorithms. We find a number of patterns relating certificate validity, host public key size, and CA signing algorithm. Figure 4 (first panel) shows that one-fourth of hosts using RSA with 2048- and 4096-bit public keys have invalid certificates. 520 government hostnames use cryptographically insecure 1024-bit RSA. In the USA, NIST issued a special public document recommending key lengths larger than 1024 with popular tools like OpenSSL being compliant [12]. We also find that RSA key sizes of 3248 bits are generally misconfigured because of incorrect usage and or 8192 bits due to lack of support in browsers for validating key sizes greater than 4096 bits. We see an increasing use of elliptic curve (EC) cryptography, dominated by 256-bit keys.

Figure 4 (second panel) shows certificate validity by signing algorithm used by the CA issuer. 920 government websites still use certificates signed using `MD5` or `SHA1` hash with RSA Encryption. A sizeable number of certificates are issued with elliptic curve (ECDSA) signatures, correlated with a higher number of valid certificates compared to RSA.

Combining these insights, in Figure 4 (third panel) we visualize the relationship between signing algorithm, public key bit size & type of the host server, and the corresponding certificate validity. Certificates are highly likely to be valid when both CAs and hosts use elliptic curve (EC) keys and signatures; *e.g.* 99% of websites

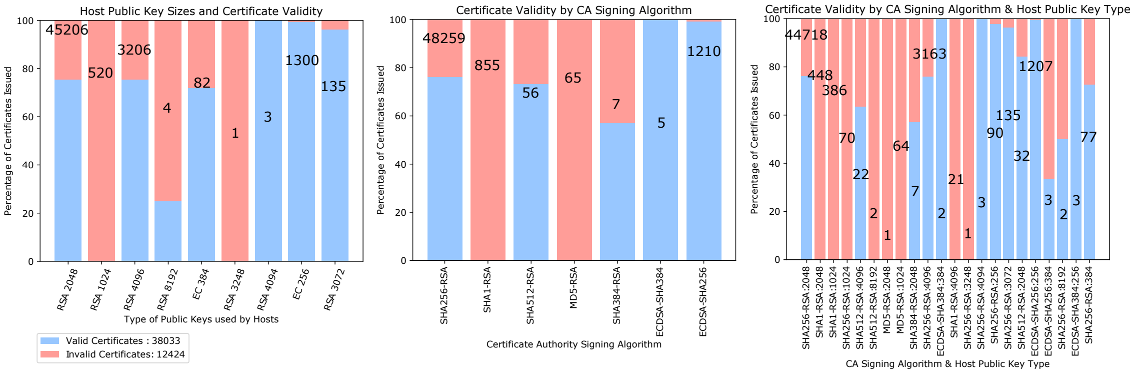

Figure 4: Worldwide: Certificate Validity/Invalidity by host cryptographic key type & CA signing algorithm. Bar colors indicate percentage of valid certificates, and the number on the bar indicates occurrences of that type.

(1200 out of 1207) where the CA signed the certificate with ECDSA-with-SHA256 attesting a 256-bit EC host public key are valid.

5.3.3 Host Public Key Pair Reuse. We notice that government websites tend to reuse wildcard certificates across different hostnames belonging to the same government, often incorrectly. One such certificate was shared across 102 hostnames in Bangladesh. However, https was invalid on *all* of these sites because of hostname mismatches; the wildcard certificate was valid for *.portal.gov.bd but was used on all *.gov.bd. In a similar case, the Colombian government used the wildcard certificate for *.micolumbiadigital.gov.co on *.gov.co. Such instances are found in 111 countries, with the top five violators being Bangladesh (2 certificates incorrectly used across 138 hostnames), Colombia (3 certificates incorrectly used across 125 hostnames), China (8 certificates incorrectly used across 107 hostnames), Dominica (1 certificate incorrectly used across 28 hostnames), and Vietnam (3 certificates incorrectly used across 21 hostnames).

Unlike cases where a single certificate is shared across different hostnames in one country, we also see instances of public key and single-certificate reuse by *different* governments. We found 58 government hostnames of 24 countries using the same certificate. 154 certificates were reused across 1,390 hostnames, with 108 certificates reused by 2 countries, 19 by 3 countries, 11 by 4 countries, and 1 by 24 countries. The most-reused certificates are invalid self-signed localhost certificates with the same set of public keys. 210 (15.1%) of these hostnames use self-signed certificates with no chain of trust, while 648 (46.6%) of the incorrectly reused ones are invalid due to hostname mismatches. This incorrect usage points to a troubling possibility that all the servers share the same private key. A malicious user with the key could observe TLS connections to a target server using the same certificate and decrypt communications with any clients who have added an exception to the invalid certificate. Valid reused certificates are wild card certificates being hosted by the same government. We do not find any instances of valid public key reuse across country governments.

5.3.4 Configuring CAA Records: DNS Certification Authority Authorization (CAA) is a DNS record type which indicate the CAs allowed to issue a certificate for the given domain. Enabling CAA records allows administrators to restrict certificate issuance to trusted CAs and implements notification procedures to identify incorrect certificate signing requests which could be initiated by masquerading adversaries [37]. We performed measurements on all of the the hostnames for CAA records and identify that only 1851 (1.36%) of all domains had a valid CAA record and 100% of the CAA records themselves were valid. The use of CAA records also mitigates the risk of certificate mis-issuance by CAs, which could have serious consequences, up to and including removal of the CA from the trust stores.

5.4 Hosting Providers

Cloud hosting providers & Content Delivery Networks (CDNs) can impact the https ecosystem by automatically enabling https without the need for customer action, sometimes for free, making them alluring for governments as an alternative to self-hosting. Some domain registrars (e.g., GoDaddy, Namecheap) also provide hosting services and simplify certificate deployment for their customers. Governments have been increasingly leveraging these platforms to streamline operations and increase their resistance to DDoS attacks. We explore the uptake of these hosting platforms and potential impacts on https adoption for their government clients (Figure 5).

We note that different countries might have different legal requirements for government cloud providers such as being FedRAMP (equivalent) certified and compliant with accessibility guidelines [36, 77]. Prior work has focused on comparing FedRAMP and the South Korean cloud certification process, presenting improvement suggestions, and analyzing improvement adoption [41, 53, 71]. These studies further motivate our case studies in Section 6.

We sort government hostnames by cloud and CDN service using the periodically updated public IP ranges published by providers like Microsoft Azure, Amazon Web Services (AWS), Cloudflare, IBM, Oracle, Google Cloud and HP-Enterprise. Akamai however does not publish an official IP range list and hence is not considered in this study. Using these CIDR prefixes, we perform lookups on the DNS A records of the domains to resolve the IP address and identify the host. We use the first IP address returned in the list of A records, and label all IP addresses not belonging to our list of service providers as "privately hosted or unknown". In Figure 5, we find

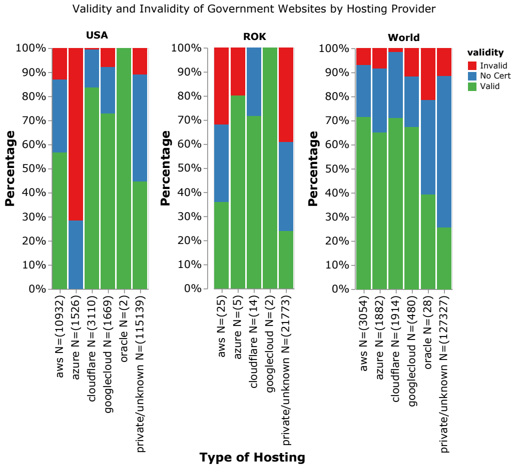

Figure 5: Certificate Validity by Hosting Type for Government Websites (case studies vs. world).

Aggregated Certificate Validity for government hostnames belonging to USA (left), ROK (center), Worldwide long tail (right)

that government websites primarily tend to be privately hosted. Those on commercial clouds or CDNs have significantly higher `https` adoption, with **60%** having valid certificates compared to **30%** on private servers.

5.5 Comparison with Non-Govt. Sites

Given the positive effects of public and commercial pressure on `https` adoption, we expected website popularity ranking and use of valid `https` to be correlated. This complicates an apples-to-apples comparison between government and non-government sites as our list includes mostly government sites outside the top millions (90.9%), for which there are no rankings. Thus, we restrict our comparison to the subset of our government hostnames present in the Tranco million dataset (12,293 of our 135,139 hostnames), comparing `https` validity while accounting for relative rank.

We compare `https` in these top government websites (mean rank: 396,427, σ: 285,611) with [1] 12,000 random, uniformly sampled top million non-government hostnames (mean rank: 499,206, σ: 286,907) and [2] 12,000 sampled top million non-government hostnames closely matching the rank distribution as the government hostnames (mean rank: 402,676, σ: 288,942). For sampling dataset [2] of non-government hostnames, we first divide the top million into (N=50) buckets by rank, and count the number of government hostnames in each bucket, ensuring each contains at least 100 government hostnames. We then uniformly sample an equal number of non-government hostnames in each bucket to match the number of government hostnames. Figure 7 compares these three sets with linear regressions on `https` validity by top million ranking, with 95% confidence interval bands.

Though ranking does have an effect, overall valid `https` use in government websites in the top million is similar to results in the long tail dataset, at ~30%. Meanwhile, the top 12,000 non-government websites have >70% valid `https` while the two non-government sets we sampled have ≈55%, indicating that even top government websites perform worse than most other top million sites. We expect that these results likely remain consistent in the long tail of the Internet. Figure 7 shows this disparity and indicates that the probability of having a valid `https` certificate reduces as ranking worsens. However, in this study we do not further compare reasons for certificate invalidity in top million non-government and government websites.

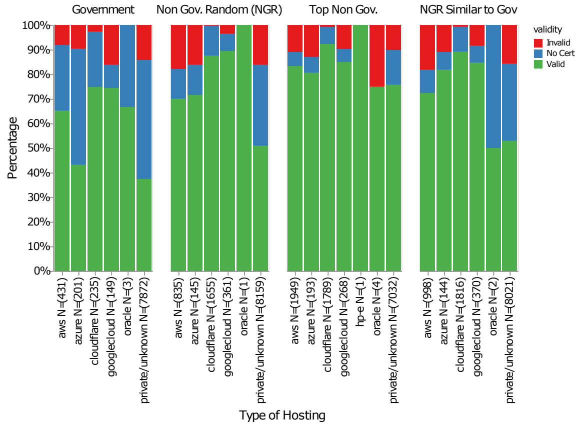

Figure 6: Certificate Validity & Hosting Type Across Non-Government vs. Government Top Millions Sites

Certificate validity by type of hosting (Cloud, CDN, and Private) for the 12K government hostnames in the Tranco top million, compared to random non-government hostnames with a rank distribution matching government ones, randomly sampled top million non-government hostnames, and the top 12K non-government hostnames.

In Figure 6, we find that privately hosted government websites in the top million have 50% of the validity of top non-government websites. In contrast, non-government websites using public cloud and CDNs have a validity greater than 70% in our sampled datasets.

6 CASE STUDIES

While the above analyses of the expanded worldwide list of government hostnames provides interesting insights, it remains unclear how representative our data sample is; for example we may have encountered only a small proportion of government hostnames from each country. To address this, we include two in-depth case studies of countries chosen because they provide public, *authoritative* lists of government hostnames: The United States of America (USA) and the Republic of Korea (ROK) or South Korea. Both governments actively curate their lists, providing a more complete view of `https` adoption. As of the publication date, both also have laws on securing government websites requiring technical measures against forgery or fraud, though only the USA's legal requirements specify `https`[60, 61, 66]. While the two countries themselves are not representative of the world, both having high human development index scores (USA:15, ROK:22) and Internet adoption rates (USA:90%, ROK:96%), among other unique factors, their relative technical sophistication likely biases them *towards* `https` adoption and thus the following analyses could be viewed as a potential high-water mark for `https` adoption among governments.

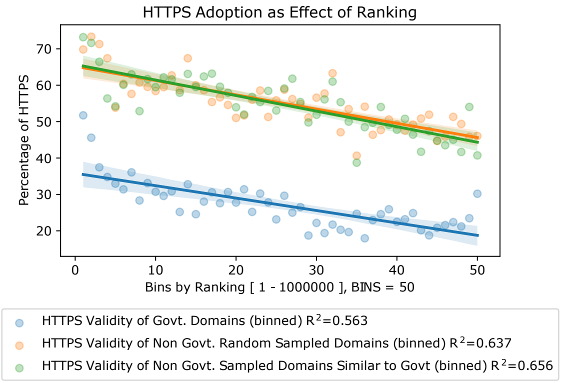

Figure 7: Valid `https` rate plotted by top million rank

Percentage of valid `https` of Government and Non Government websites in the Tranco top million by ranking, with sites grouped into 50 bins. Plotted data are the 12K government hostnames listed (blue), randomly sampled non-government hostnames (orange), and randomly sampled non-government hostnames with a rank distribution matching the government sites (green). Linear models show a trend of decreasing `https` with rank for all sets, with worse `https` adoption for government sites.

6.1 Case Study 1: United States of America

The United States General Services Administration (GSA) publishes an open data set of all hostnames which belong to the government and archives this information for each presidential term [78, 79]. We consider this data the ground truth set for the United States and perform the same analyses as we do with the worldwide lists. The GSA categorizes the hostnames into federal, state, local, regional, county, native sovereign nations and quasi-governmental hostnames. We additionally merge this with the public list of military hostnames obtained from the Department of Defense (DoD). While the USA predominantly uses `.gov` as the official government extension, it also uses other domain extensions such as `.fed.us` and `.mil` for its federal and military related websites. As mentioned in section 4.1.1, we filtered only for hostnames with official government domain extensions (such as `.gov` in the case of the US), excluding the others. From the authoritative GSA list, US government websites demonstrate overall **81.12%** valid `https` use.

6.1.1 Certificate Analysis. Let's Encrypt is the most popular CA used by US government websites and less than 5% of the Let's Encrypt certificates used by these websites are invalid as shown in Figure 8. We identify that 83.11% of the invalid Let's Encrypt certificates are invalid due to hostname mismatches, 13.39% due to expiry of the certificate and the remaining due to the client being unable to validate the root certificate in the local trust store.

Consistent with the global scan, we find that CAs using elliptic curves for signatures tend to be valid & used correctly. Figure 9 shows certificate validity by type of signing algorithm used by the CA authority and public key sizes used by the host. We observe that 100% of certificates are invalid when issued using `SHA1` with RSA encryption, `MD5` with RSA encryption or Probabilistic Signature Scheme (PSS), whereas 100% of certificates are valid using ECDSA with `SHA384`. We also find that valid certificates are highly clustered together in duration of validity, within 3 years of the current date (see Figure 10), compared to invalid certificates which are issued for much longer durations than agreed upon by the CAs [13, 14].

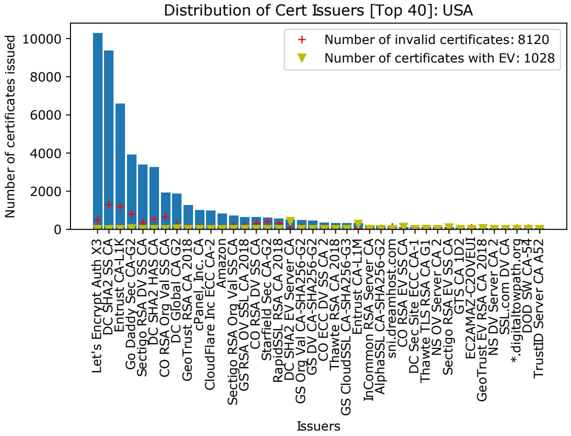

Figure 8: USA Case: Certificate Validity by Issuing Authority

Abbr: COMODO=CO, Network Solutions=NS, DigiCert=DC, GlobalSign=GS, GlobalTrust=GT, Encryption Everywhere=EE, "High Assurance Server"=HAS, "Secure Server"=SS

6.1.2 Hosting Analysis. Government web hosting both worldwide and in the US is dominated by private hosting. 13.02% of US government sites are on the public cloud and CDNs, close to the 11.46% for all government sites in the Tranco top million. Government sites are 3.5x more likely to be hosted on AWS than the second most popular service Cloudflare, with Azure and Google Cloud closely following, as shown in Figure 5. The public GSA dataset contains a large number of *unreachable* sites in the 2016 presidential end-of-term snapshot. This is because these websites are either archived or unavailable. We do not consider unavailable websites in our analyses. In Appendix A.1, we discuss validity by host for the individual datasets aggregated here.

6.2 Case Study 2: South Korea

The government of South Korea (ROK) centrally maintains a search portal under "gov.kr", also called "Government24," which contains a comprehensive organizational map of the Korean government and serves as an authoritative database of all of their hostnames [3]. All 21,885 hostnames present in the database at the time were scraped from the search results and de-duplicated. From this authoritative government website list, we measured a **37.95%** rate of `https` validity in the Republic of Korea.

6.2.1 Certificate Analysis. The largest issuer of government web certificates in South Korea was the CA Sectigo RSA DV Secure Server, closely followed by Alpha SSL (see Figure 11)–also the 4th and 10th largest CAs used worldwide (Figure 2), and the 5th and 24th largest CAs used by US government websites (Figure 9). 47% of ROK certificates issued by Sectigo RSA DV were invalid, due to hostname mismatches, inability to get local issuer certificate, a self-signed certificate in the chain, or expiry. ROK continues to use private CAs or CAs which were previously a part of the

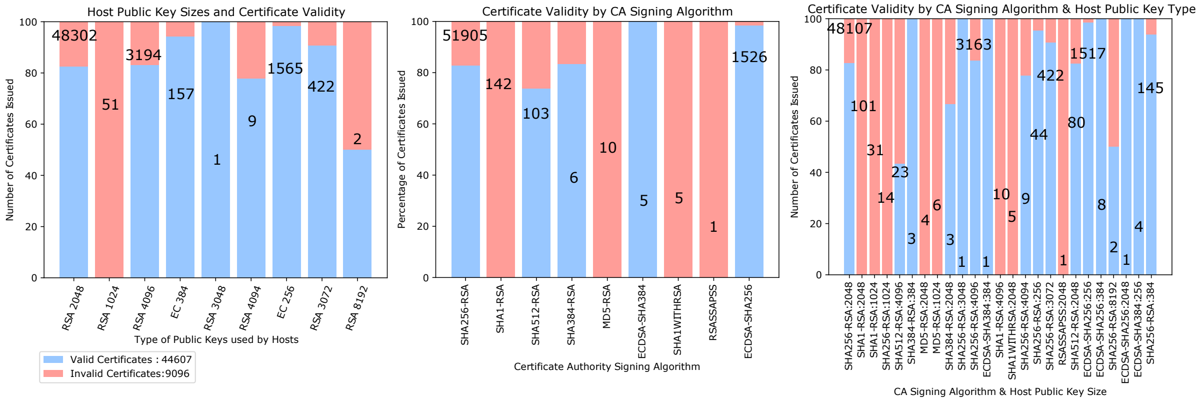

Figure 9: USA Case: Certificate validity by key type and CA signing algorithm for government domains.

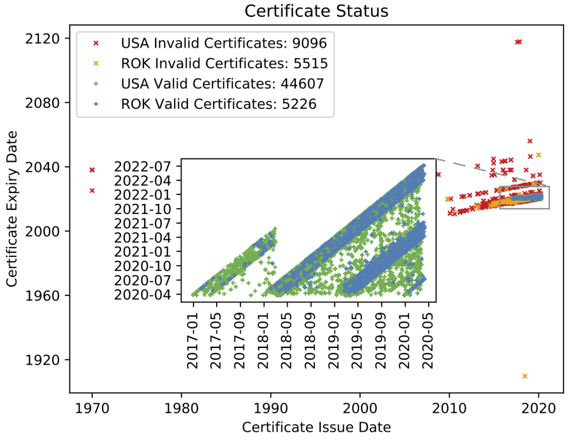

Figure 10: USA & ROK: Certificate Validity by Issue Date

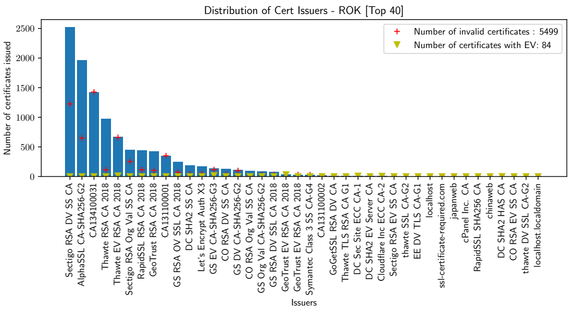

Figure 11: ROK Case: Cert Validity by Issuing Authority

Abbr: COMODO=CO, DigiCert=DC, GlobalSign=GS, GlobalTrust=GT, Encryption Everywhere=EE, "High Assurance Server"=HAS, "Secure Server"=SS

NPKI infrastructure and are now untrusted by all major browsers and operating systems. Our results of the breakdown of certificate validity by cryptographic key and algorithm usage in Figure 12 indicates a higher validity for elliptic curves.

6.2.2 Hosting Analysis. As shown in Figure 5, the vast majority of government websites in the ROK are privately hosted, with only 0.21% of sites being hosted on popular large public clouds or CDNs. Similar to the USA, as shown in Figure 10, we notice the valid certificates being clustered together.

6.3 Case Study Discussion

Despite having similar human development index scores and Internet adoption rates, overall `https` adoption by the US and South Korean governments are very different (at 81.12% and 37.95%, respectively). `https` error and misconfiguration profiles also differ between the countries. Exceptions, such as unsupported SSL protocol, timeout, connection refused or reset by peers, and wrong SSL version number, compose 2.79% of `https` invalidity in the USA, as compared to 21.08% in the ROK. In the US, the usage of a self-signed certificate in the certificate chain causes 0.18% of errors and the inability to get local issuer certificate composes 2.44%. In the ROK, the corresponding figures are 5.95% and 15.44%, respectively.

South Korea created and deployed its own CA accredited by the National Public Key Infrastructure (NPKI), which historically was only accessible through a plugin installed by the citizens to ensure secure access to government resources with a user-issued identity certificate [43]. Over the past decade, there have been massive efforts to improve interoperation between NPKI and web standardization efforts by the W3C and EFF for PKI [50, 65]. In 2018, South Korea prepared a bill to abolish the government-accredited NPKI and switch to the web standard [43]. Since then, the recent Electronic Government Act and proceeding Enforcement Decree have required (without precisely defining) security measures against forgery or data theft for e-government services [60, 61], underscoring the importance of scanning government websites for accountability. Two years later (Figure 11) we continue to see NPKI-attested sub CAs, such as `CA134100031` (3rd most popular) and `CA131100001` (9th most popular), being used by government websites in the ROK, but treated as invalid certificates by popular browsers and tools like OpenSSL due to their repeated violation of certificate issuance standards [15].

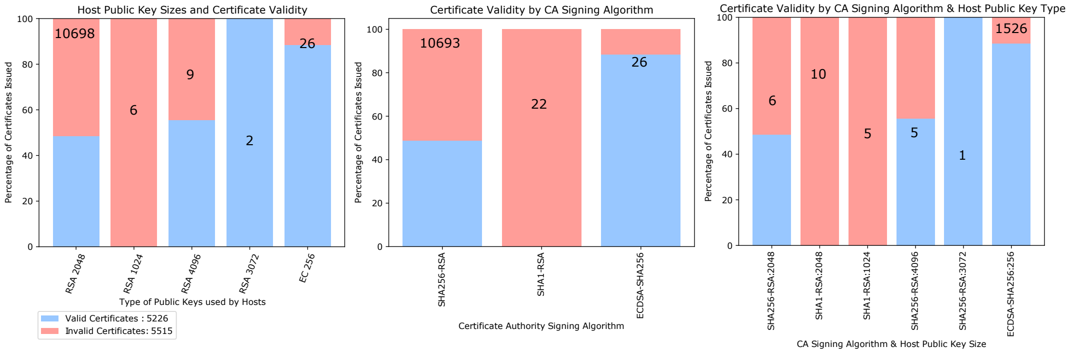

Figure 12: South Korea Case: Certificate validity by key type and CA signing algorithm for government domains.

7 DISCUSSION

In the remainder of the paper, we discuss the limitations of our work (section 7.1), our disclosure process, and the effectiveness of the disclosures (section 7.2). We further present our experiential perspective on the importance of https for government websites with examples (section 7.3), and conclude by providing our recommendations to improve https adoption for governments (section 8).

7.1 Study Limitations

7.1.1 Biases. As mentioned in Section 4.1.1, the study conservatively selects only hostnames with a valid government TLD like .gov, .gouv, .gub, .guv, .gob, .go followed by valid country codes, excepting the hand-curated list. However, we note that some governments do use other TLDs, like .net, .org or .com, for their official websites. Such websites, even with valid https, would largely be excluded from this study.

Additionally, the measurement is a single snapshot of the state of https adoption in governments across the world and not a longitudinal study. Our results do not account for natural churn in website availability and https support, on which we do not make claims, though it may be interesting in future work to document “gaps” in https for important websites. We also acknowledge country-based biases in our dataset, towards large countries with higher technology adoption like the USA. While our work aims to combat these biases by seeking the “long tail” of websites, it fails to avoid them completely. Such biases are inherent in the Internet itself; recent estimates by Solarwinds indicate that the United States hosts approximately 43% of the world’s top million websites followed by Germany and China respectively [67]. This skew may persist for government websites; for example, the USA has 6 times as many reachable government websites as South Korea. Larger or more populous countries may have more websites split at the regional and county levels compared to smaller countries.

7.1.2 Location. We perform this snapshot from a single location instead of using geographically distributed scanners; censorship by a country’s firewall could affect our snapshot. For example, the “Great Firewall of China” [1] has been particularly challenging. We were only able to reach around 50% of Chinese government hostnames in our crawls, and around 20% of Chinese hostnames in the top millions seed list. Figure 1 shows a very low rate of valid https among reachable sites for China. Using a VPN service (provided through ExpressVPN [28]) to crawl from locations closer to China, such as Hong Kong, did not yield better results. Lack of access to most Chinese websites weakens our claims about https in China, but we believe our estimate is not wildly inaccurate as we have still scanned 22,487 Chinese sites. (Of these sites, 13,080 use https while 9,407 use http. Of those using https, only 11% (1,438) use it correctly, with similar reasons for invalidity as worldwide. 60.1% (7,861) of cases are invalid because of hostname mismatches, and 16.23% (2,124) are invalid because of the inability to get a local issuer certificate. 9.68% (1,267) websites use a self-signed certificate, 0.4% use a self-signed certificate in the chain, and 2.56% use an expired certificate.) Given the difficulty of travel to China in recent months and administrative barriers to accessing local datacenters and cloud providers, it seems unlikely in the short term that we will get better measurements for these sites.

7.1.3 Sampling. Our comparisons of government vs. non government websites in section 5.5 used a set of randomly sampled non-government websites in the top millions with a similar ranking distribution to the government sites. While we believe our analysis after sampling is accurate, scanning the whole top million non-government websites instead of sampling could increase our result’s precision. The effects of rank presented in Figure 7 also implicitly assume that the 20,000 sites in each bin are similar; https adoption for each bin is measured with only a statistically significant representative sample.

7.1.4 Alternatives. Finally, other countries such as India, United Kingdom, and Australia provide authoritative listings of government websites and could have been chosen as case studies. We chose the USA and South Korea for their similarity in Internet adoption rates and human development index scores across language and culture differences, but a larger set of cases and a longitudinal analysis of the government websites would provide more nuanced views.

7.2 Notification & Disclosure

Government domain name registrations are typically handled by a separate registrar for each government and expected to meet stringent verification requirements. For example, on March 5, 2020 the US (dotgov) registrar made it mandatory to obtain notarized signatures on authorization letters when requesting a .gov hostname [35] due to recent attacks [45] that allowed an impersonator to register a government hostname.

As a part of our analysis, we generated reports per country of potentially vulnerable hostnames, including invalid `https`, failed upgrades of `http` to `https`, and unreachable hostnames which were still linked from other pages that we discovered during the analysis. We emailed the respective countries' government domain registrars (performing *whois* queries on the country registrars to find listed technical contacts), included their vulnerable hostnames as a file attachment and requested contact information for the domain owners or appropriate forwarding of our reports. We sent emails to 182 countries, since 9 (Angola, Benin, Democratic Republic of Congo, Estonia, Guinea, Netherlands, Norway, Switzerland, and Vanatu) countries had `https` for every detected hostname, and at the time of disclosure we still had no hostnames for 7 countries due to our conservative filtering. 175 emails were delivered, and 7 bounced. We retried the 7 countries by emailing the listed administrative contact, of which 4 emails failed again. 6 registrars sent an automated message acknowledging the receipt of our email. Some of the countries with valid `https` for all hostnames had very few in total (≈30); an in-depth analysis of these countries may be needed to clarify the correctness of this result.

Responses were surprisingly positive. 39 domain registrars were supportive: 3 (Brazil, Lebanon, and Liberia) provided us with the necessary contact information, 13 of them (Austria, Bosnia & Herzegovina, Burundi, Cayman Islands, Columbia, Lithuania, Netherlands, Nigeria, Rwanda, Sri Lanka, Tanzania, Tonga, and Ukraine) re-directed our emails to the corresponding government authority or a responsible person who was the intended recipient, and 2 (Japan, and Norway) responded mentioning that they could not provide us with contact information that isn't publicly available in the `whois` and suggested that we query their *whois* servers for the information. The registrar of one country responded negatively, saying *"We are not interested"*.

Previous work by Stock *et al.* showed that transmission of vulnerability reports to actual domain owners had very limited impact on successful resolution, and only resulted in a very small (∼5.8%) number of emails being actually received [75]. In our study, 22% of the country government domain registrars / CERT authorities proactively replied to our messages and have begun taking the necessary steps to fix their certificates. We believe that for government domains, the registrar, who might represent a government body, may have a higher incentive to respond to such reports than individual developers and is in a position of power to make a meaningful change. In section 7.2.2 we discuss a follow-up scan to measure the effectiveness of our notifications, conducted 2 months after.

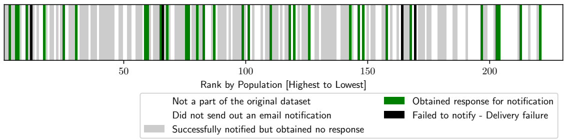

Figure 13: Response by Country Population (Rank)

7.2.1 Experiences with Disclosure to Governments. Interestingly, we found a pattern where domain registrars of countries with the highest population were least communicative and responsive to our emails, but saw a much higher response rate from countries with medium or low populations as shown by the higher density of green stripes between rank 50-100 and after 200 in the center of Figure 13. 34 countries are territories of other countries (white bands in the figure), and thus were not included in our study. Despite the lack of responses from other countries which were successfully notified, we notice possible silent updates to the websites addressing the issue as presented in section 7.2.2.

7.2.2 Notification Effectiveness. Two months after notification, we scanned the 15,179 government websites with previously invalid `https` to understand notification effectiveness. 1,572 of these were unreachable and seem to have been removed, while 1,263 websites had fixed the certificate invalidity issues. 12,344 sites continue to serve content with invalid certificates. Assuming that newly unavailable websites (no longer returning a 200 status) have been removed on purpose by webmasters and considering this a fix, we optimistically estimate improvement at 18.7%; otherwise, the improvement is only **8.3%**.

Of the 47,458 sites unreachable in our original scan and thus not considered in our list of 135,139 sites, we notice that 38,077 continue to be unreachable, while 2,850 (6%) sites now serve content using an invalid certificate, and 6,531 (**13.76%**) sites serve content with a valid certificate. 950 (**1.15%**) websites which previously served http-only traffic now serve valid `https` traffic, while 1,523 (1.85%) websites serve content with an invalid certificate, and the remaining (96.9%) continue to use only `http`.

Preliminary findings indicate that notifications and disclosure do have a small positive effect, with 62 countries showing at least a 10% improvement in valid `https` and 7 countries (Bahrain, Burkina Faso, Cuba, Honduras, Portugal, Libya, and Vietnam) showing improvement above 40%. Since we do not perform a full re-scan of all hostnames and only measure changes in sites which were previously invalid, we cannot measure deterioration of websites. The United States government issued a statement after our disclosure mandating HSTS preloading for ".gov" websites by September 1, 2020 [21]. However, we cannot definitively attribute positive changes and the improvements to our disclosures.

7.3 Why Should Governments Care?

We discuss a number of specific threats due to lack of valid `https` and risky certificate infrastructure.

7.3.1 Censorship. Without `https`, other countries' governments may censor content for their nation. In 2015, Russia's efforts to censor specific content on Wikipedia failed due to its use of `https`, and they were left with the choice to either ban the platform completely or let the content be accessible uncensored [32].

7.3.2 Attacks with Valid Certificates. Prior work has also demonstrated *compelled certificate creation attacks*, in which government agencies can compel a CA to issue false certificates, which can then be used to intercept secure communication for surveillance purposes [72]. A disproportionate number of CAs are US organizations, potentially vulnerable to compulsion by the US government. For example, the Mozilla NSS trusted CA store indicates 42 CAs registered in the USA, followed in second place by Bermuda and Spain with 6 CAs each, 4 each in Taiwan, China, India and Belgium. The USA is home to 7 times more trusted root CAs than the next country in which CAs are based. Any CAs registered in other countries are also vulnerable to their respective governments.

Automated certificate issuance has become extremely simple and almost zero cost to the domain owner due to services like Let's Encrypt. Attackers can purchase domain names mimicking government websites (see Section 4.1.1) from private registrars such as GoDaddy, Namecheap, *etc.*, albeit for a higher price (approx. 150 USD). For example, we found a website registered with the Sierra Leone country code, etagov.sl, posing as the Sri Lankan government's travel authorization portal eta.gov.lk, with a valid certificate. Uninformed users might mistake the country code .sl for Sri Lanka, when in fact it is .lk and register for a visa on the phishing website resulting in identity theft. We have responsibly disclosed this vulnerability to the LK Domain Registrar, Lanka Government Information Infrastructure. We also find 85 unique hostnames which end with "gov.us" of the format abcgov.us, indicating the very real threat of carrying out such spoofing attacks with perfectly valid certificates obtained from a free CA or CDN.

7.3.3 Cross-Government Links. Finally, in our crawls we noticed a lot of cross-government links between different countries. Austrian government sites contain the largest number, linking to 70 other governments. Such links should be carefully curated, as they could be used to find and exploit vulnerable government sites. Links not using https could misdirect users to MITM versions, spreading misinformation, hate speech, or carrying out phishing attacks on visitors from specific other countries.

8 RECOMMENDATIONS

Based on our results, we outline possible security threats and recommend improvements to secure government sites. While these recommendations may involve challenging infrastructural changes, we believe they can be tractable with the right industry, government, and open-source collaborations, and would strongly improve the certificate ecosystem and security of government websites.

8.1 For Certificate Authorities

Currently, Let's Encrypt issues challenges to domain owners to provision a DNS record or an http resource under a known URI for the requested domain, thereby binding public keys to the domain. In response to the problem of public key reuse (Section 5.3.3), we recommend that CAs perform additional checks to see if the public key has already been issued a certificate for another hostname and check if the new hostname is a sub-domain of the previous one. Coupled with certificate transparency, this could make it easier to discover hosts using compromised or repeated keys.

CAs like Let's Encrypt might also be able to selectively issue EV certificates for government hostnames based on a digital signature on government domain information from the country CERT division, or collaborate with different governments' root CA authorities to provision a certificate with the country CA certificate as an intermediate in the chain. The inclusion of an intermediate government CA provides the ability for local governments to validate the signing requests, and coupled with the CAA records, could prevent distrust with other governments where the CA services operate from or whose legal regulations they should comply with.

8.2 For Domain Registrars and Owners

Given the possibility of domain-based spoofing or phishing attacks, domain registrars should pay special attention to domain names including key words involving government functions and potentially establish special checks.

In conducting our study, we found that most countries have a separate CERT division or cybersecurity response center with a country-specific root CA, often used for internal purposes such as citizen identity, digital signatures, *etc.* We suggest using country government CAs as intermediate CAs for government websites, with the root CA as Let's Encrypt or another CA. As shown, domain owners and webmasters operating government sites tend to use Let's Encrypt due to ease. While this may be better than having no certificate, other options exist, albeit with increased bureaucratic barriers, which could be circumvented with collaboration between CAs and government root CAs. Additionally, we recommend governments include DNSSEC signed CAA records for their websites so that only trusted CAs can issue certificates. We also recommend that domain owners enlist government websites into the HTTP Strict Transport Security (HSTS) Preload list directing browsers to always use TLS to communicate with the website [39].

9 CONCLUSION

Many (≈72%) government sites do not still use https, either due to lack of TLS infrastructure or a large variety of certificate errors. Through our study of 135,139 government websites across the Internet from almost every country in the world, we have identified major categories and frequencies of these errors, with an aim to privately disclose to the sites' web administrators and measure the effects. Common errors include the misconfiguration of hostnames, expired certificates, reuse of keys and serial numbers between sites likely to be hosted on different servers or use of default certificates, using insecure cryptographic algorithms such as MD5, or SHA1, certificate self-signing, and other issues. In the US, our recommendations can be used to improve compliance with the DOTGOV Online Trust in Government Act of 2019 outlining requirements (including https) for the .gov domain, facilitating the technical security practices needed to maintain public trust in the government [66].

ACKNOWLEDGEMENTS

We would like to thank our shepherd Matthew Luckie and the anonymous reviewers for their valuable feedback which shaped the final paper. We also thank Dan Ports, Ming Liu, and the UW CSE Support team for their help in accessing the infrastructure to run the measurements. For their valuable feedback and discussions, we thank Chris Thompson from Google, and Matt Johnson and Spencer Sevilla from the ICTD Lab. We thank Tae Oon Jang for his knowledge and help in navigating Korean e-government resources.

REFERENCES

[1] [n.d.]. The Great Firewall of China — web of control | Financial Times. https://www.ft.com/content/e19b3022-40eb-11e9-9bee-efab61506f44. (Accessed on 03/06/2020).

[2] 2016. 95% of HTTPS servers vulnerable to trivial MITM attacks | Netcraft News. https://news.netcraft.com/archives/2016/03/17/95-of-https-servers-vulnerable-to-trivial-mitm-attacks.html. (Accessed on 09/11/2020).

[3] 2020. Government/Local Government Operation Site | Agency Information | Government 24. https://www.gov.kr/portal/orgSite?. (Accessed on 09/12/2020).

[4] 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* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery, New York, NY, USA, 2473–2487. https://doi.org/10.1145/3319535.3363192

[5] Mustafa Emre Acer, Emily Stark, Adrienne Porter Felt, Sascha Fahl, Radhika Bhargava, Bhanu Dev, Matt Braithwaite, Ryan Sleevi, and Parisa Tabriz. 2017. Where the Wild Warnings Are: Root Causes of Chrome HTTPS Certificate Errors. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17)*. Association for Computing Machinery, New York, NY, USA, 1407–1420. https://doi.org/10.1145/3133956.3134007

[6] Devdatta Akhawe, Bernhard Amann, Matthias Vallentin, and Robin Sommer. 2013. Here's My Cert, so Trust Me, Maybe? Understanding TLS Errors on the Web. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 59–70. https://doi.org/10.1145/2488388.2488395

[7] Johanna Amann, Oliver Gasser, Quirin Scheitle, Lexi Brent, Georg Carle, and Ralph Holz. 2017. Mission accomplished? HTTPS security after DigiNotar. In *Proceedings of the 2017 Internet Measurement Conference*. 325–340.

[8] Jacob Appelbaum. 2011. Detecting certificate authority compromises and web browser collusion. *Tor Blog* 22 (2011).

[9] Apple. 2018. List of available trusted root certificates in MacOS. https://support.apple.com/en-gb/HT208127. (Accessed on 06/02/2020).

[10] Cory L Armstrong. 2011. Providing a clearer view: An examination of transparency on local government websites. *Government Information Quarterly* 28, 1 (2011), 11–16.

[11] Zineb Ait Bahajji and Gary Illyes. 2014. Official Google Webmaster Central Blog: HTTPS as a ranking signal. https://webmasters.googleblog.com/2014/08/https-as-ranking-signal.html. (Accessed on 03/03/2020).

[12] Elaine Barker and Allen Roginsky. 2015. SP 800-131 (June 11 2010): Recommendation for the Transitioning of Cryptographic Algorithms and Key Lengths. https://csrc.nist.gov/csrc/media/publications/sp/800-131/archive/2010-06-16/documents/sp800-131-draft2-june2010.pdf. (Accessed on 03/06/2020).

[13] CA Browser Forum. 2017. Ballot 193 - 825-day Certificate Lifetimes - CAB Forum. https://cabforum.org/2017/03/17/ballot-193-825-day-certificate-lifetimes/. (Accessed on 05/24/2020).

[14] CA Browser Forum. 2019. Ballot SC22 - Reduce Certificate Lifetimes (v2) - CAB Forum. https://cabforum.org/2019/09/10/ballot-sc22-reduce-certificate-lifetimes-v2/. (Accessed on 05/24/2020).

[15] Chromium. 2018. 823665 - please remove trust of GPKIRootCA1 root certificate or sub-ca - chromium. https://bugs.chromium.org/p/chromium/issues/detail?id=823665. (Accessed on 05/29/2020).

[16] Cisco. 2016. Umbrella Popularity List. (Accessed on 09/07/2020).

[17] Jeremy Clark and Paul C Van Oorschot. 2013. SoK: SSL and HTTPS: Revisiting past challenges and evaluating certificate trust model enhancements. In *2013 IEEE Symposium on Security and Privacy*. IEEE, 511–525.

[18] Cloudflare. [n.d.]. What Happens in a TLS Handshake? | SSL Handshake | Cloudflare. https://www.cloudflare.com/learning/ssl/what-happens-in-a-tls-handshake/. (Accessed on 05/24/2020).

[19] Cloudflare. 2019. CFSSL: Cloudflare's PKI and TLS toolkit. https://github.com/cloudflare/cfssl.

[20] OWASP Contributors. 2020. Man-in-the-middle Software Attack | OWASP Foundation. https://owasp.org/www-community/attacks/Man-in-the-middle_attack. (Accessed on 09/11/2020).

[21] DotGov. 2020. Making .gov More Secure by Default | DotGov. https://home.dotgov.gov/management/preloading/dotgovhttps/. (Accessed on 09/12/2020).

[22] Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, and J. Alex Halderman. 2015. A Search Engine Backed by Internet-Wide Scanning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. Association for Computing Machinery, New York, NY, USA, 542–553. https://doi.org/10.1145/2810103.2813703

[23] Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, and J. Alex Halderman. 2015. A Search Engine Backed by Internet-Wide Scanning. In *22nd ACM Conference on Computer and Communications Security*.

[24] Zakir Durumeric, David Adrian, Ariana Mirian, James Kasten, Elie Bursztein, Nicolas Lidzborski, Kurt Thomas, Vijay Eranti, Michael Bailey, and J. Alex Halderman. 2015. Neither Snow Nor Rain Nor MITM...: An Empirical Analysis of Email Delivery Security. In *Proceedings of the 2015 Internet Measurement Conference* (Tokyo, Japan) *(IMC '15)*. Association for Computing Machinery, New York, NY, USA, 27–39. https://doi.org/10.1145/2815675.2815695

[25] Zakir Durumeric, James Kasten, Michael Bailey, and J Alex Halderman. 2013. Analysis of the HTTPS certificate ecosystem. In *Proceedings of the 2013 conference on Internet measurement conference*. 291–304.

[26] Zakir Durumeric, Zane Ma, Drew Springall, Richard Barnes, Nick Sullivan, Elie Bursztein, Michael Bailey, J Alex Halderman, and Vern Paxson. 2017. The Security Impact of HTTPS Interception.. In *NDSS*.

[27] Electronic Frontier Foundation EFF. [n.d.]. Certbot. https://certbot.eff.org/. (Accessed on 09/12/2020).

[28] ExpressVPN. [n.d.]. High-Speed, Secure & Anonymous VPN Service | ExpressVPN. https://www.expressvpn.com/. (Accessed on 06/02/2020).

[29] Tariq Fadai, Sebastian Schrittwieser, Peter Kieseberg, and Martin Mulazzani. 2015. Trust me, I'm a Root CA! Analyzing SSL Root CAs in Modern Browsers and Operating Systems. In *2015 10th International Conference on Availability, Reliability and Security*. IEEE, 174–179.

[30] Sascha Fahl, Yasemin Acar, Henning Perl, and Matthew Smith. 2014. Why Eve and Mallory (Also) Love Webmasters: A Study on the Root Causes of SSL Misconfigurations. In *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security* (Kyoto, Japan) *(ASIA CCS '14)*. Association for Computing Machinery, New York, NY, USA, 507–512. https://doi.org/10.1145/2590296.2590341

[31] Adrienne Porter Felt, Richard Barnes, April King, Chris Palmer, Chris Bentzel, and Parisa Tabriz. 2017. Measuring {HTTPS} Adoption on the Web. In *26th {USENIX} Security Symposium ({USENIX} Security 17)*. 1323–1338.

[32] Electronic Frontier Foundation. 2015. Russia's Wikipedia Ban Buckles Under HTTPS Encryption. https://www.eff.org/deeplinks/2015/08/russias-wikipedia-ban-buckles-under-https-encryption. (Accessed on 03/12/2020).

[33] Patrick Nohe (GlobalSign). 2020. One Year Certificates - Maximum SSL/TLS Certificate Validity is Now One Year. https://casecurity.org/2020/07/09/one-year-certs/. (Accessed on 09/07/2020).

[34] GoDaddy. 2019. EV SSL Certificate | Buy an Extended Validation Certificate - GoDaddy. https://www.godaddy.com/web-security/ev-ssl-certificate. (Accessed on 09/21/2020).

[35] United States Dot Gov. 2020. Recent Updates. https://home.dotgov.gov/. (Accessed on 05/25/2020).

[36] Melvin Greer. 2015. FITARA and FedRAMP: Accelerating federal cloud adoption. *IEEE Cloud Computing* 2, 5 (2015), 48–52.

[37] Phillip Hallam-Baker, Rob Stradling, and B Laurie. 2013. DNS certification authority authorization (CAA) resource record. *Internet Engineering Task Force* (2013), 6844.

[38] Nadia Heninger, Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. 2012. Mining Your Ps and Qs: Detection of Widespread Weak Keys in Network Devices. In *Presented as part of the 21st USENIX Security Symposium (USENIX Security 12)*. USENIX, Bellevue, WA, 205–220. https://www.usenix.org/conference/usenixsecurity12/technical-sessions/presentation/heninger

[39] Jeff Hodges, Collin Jackson, and Adam Barth. 2012. Http strict transport security (hsts). (2012).

[40] ICANN. 2017. Country code top-level domain - ICANNWiki. https://icannwiki.org/Country_code_top-level_domain. (Accessed on 03/06/2020).

[41] Hangoo Jeon and Kwang-Kyu Seo. 2015. A framework and improvements of the Korea cloud services certification system. *The Scientific World Journal* 2015 (2015).

[42] Dixon Jones. 2012. Majestic Million CSV now free for all, daily. (Accessed on 09/07/2020).

[43] Keechang Kim. 2018. Woes of government driven 'standard' – Korean PKI implementation 1999-2018. https://www.standardsuniversity.org/e-magazine/october-2018-volume-9-issue-3-privacy-freedom-human-rights/woes-of-government-driven-standard-korean-pki-implementation-1999-2018/. (Accessed on 05/29/2020).

[44] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[45] Brian Krebs. 2019. It's Way Too Easy to Get a .gov Domain Name — Krebs on Security. https://krebsonsecurity.com/2019/11/its-way-too-easy-to-get-a-gov-domain-name/. (Accessed on 05/25/2020).

[46] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. 2017. " I Have No Idea What I'm Doing"-On the Usability of Deploying {HTTPS}. In *26th {USENIX} Security Symposium ({USENIX} Security 17)*. 1339–1356.

[47] Adam Langley. 2012. SSL interstitial bypass rates.

[48] Ben Laurie and Cory Doctorow. 2012. Secure the internet. , 325–326 pages.

[49] 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

[50] Seok-Lae Lee, Jae-il Lee, and Hong-Sub Lee. 2001. Global PKI Interoperability: Korean Endeavour. In *Proc. of the first International Workshop for Asian PKI, Korea*. Citeseer.

[51] Antonis Manousis, Roy Ragsdale, Ben Draffin, Adwiteeya Agrawal, and Vyas Sekar. 2016. Shedding light on the adoption of let's encrypt. *arXiv preprint arXiv:1611.00469* (2016).

[52] Stephanos Matsumoto and Raphael M Reischuk. 2015. Certificates-as-an-Insurance: Incentivizing accountability in SSL/TLS. In *Proceedings of the NDSS Workshop on Security of Emerging Network Technologies (SENT'15)*.

[53] Kevin McGillivray. 2015. FedRAMP, Contracts, and the US Federal Government's Move to Cloud Computing: If an 800-Pound Gorilla Can't Tame the Cloud, Who Can. *Colum. Sci. & Tech. L. Rev.* 17 (2015), 336.

[54] Microsoft. 2019. List of Participants - Microsoft Trusted Root Program. https://docs.microsoft.com/en-us/security/trusted-root/participants-list. (Accessed on 09/07/2020.

[55] Ariana Mirian, Christopher Thompson, Stefan Savage, Geoffrey M Voelker, and Adrienne Porter Felt. 2018. HTTPS Adoption in the Longtail. (2018).

[56] Bodo Möller, Thai Duong, and Krzysztof Kotowicz. 2014. This POODLE bites: exploiting the SSL 3.0 fallback. *Security Advisory* (2014).

[57] Mozilla. 2017. Mozilla Included CA Certificate List. https://wiki.mozilla.org/CA/Included_Certificates. (Accessed on 09/07/2020).

[58] Mozilla. 2020. Cert Verifier - Extended Validation. https://hg.mozilla.org/mozilla-central/file/tip/security/certverifier/ExtendedValidation.cpp. (Accessed on 03/04/2020).

[59] David Naylor, Alessandro Finamore, Ilias Leontiadis, Yan Grunenberger, Marco Mellia, Maurizio Munafò, Konstantina Papagiannaki, and Peter Steenkiste. 2014. The cost of the" s" in https. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*. 133–140.

[60] Ministry of Public Administration and Security. [n.d.]. National Law Information Center | Law> Text-e-Government Law. http://law.go.kr/LSW/lsInfoP.do?lsiSeq=213795&ancYd=&ancNo=&efYd=20200805&nwJoYnInfo=Y&ancYnChk=0&efGubun=Y&vSct=%EC%A0%84%EC%9E%90%EC%A0%95%EB%B6%80%EB%B2%95#0000.

[61] Ministry of Public Administration and Security. [n.d.]. National Law Information Center | Law> Text-Enforcement Decree of the e-Government Act. http://law.go.kr/LSW/lsInfoP.do?lsiSeq=206157&ancYd=&ancNo=&efYd=20190101&nwJoYnInfo=Y&ancYnChk=0&efGubun=Y&vSct=%EC%A0%84%EC%9E%90%EC%A0%95%EB%B6%80%EB%B2%95#0000.

[62] OpenSSL. [n.d.]. OpenSSL: Cryptography and SSL/TLS toolkit. https://www.openssl.org/. (Accessed on 03/06/2020).

[63] OpenSSL. [n.d.]. Verify - OpenSSL. https://www.openssl.org/docs/man1.0.2/man1/verify.html. (Accessed on 09/21/2020).

[64] Gustaf Ouvrier, Michel Laterman, Martin Arlitt, and Niklas Carlsson. 2017. Characterizing the HTTPS trust landscape: a passive view from the edge. *IEEE Communications Magazine* 55, 7 (2017), 36–42.

[65] Hun Myoung Park. 2012. The web accessibility crisis of the Korea's electronic government: Fatal consequences of the digital signature law and public key certificate. In *2012 45th Hawaii International Conference on System Sciences*. IEEE, 2319–2328.

[66] Sen. Gary C Peters. 2019. DOTGOV Online Trust in Government Act of 2019. https://www.congress.gov/bill/116th-congress/senate-bill/2749/text.

[67] Solarwinds Pingdom. 2012. The US hosts 43% of the world's top 1 million websites. https://www.pingdom.com/blog/united-states-hosts-43-percent-worlds-top-1-million-websites/. (Accessed on 09/12/2020).

[68] Yaroslava Ryabova. 2018. HTTPS does not mean a site is safe | Kaspersky official blog. https://www.kaspersky.com/blog/https-does-not-mean-safe/20725/. (Accessed on 03/12/2020).

[69] Kazunori Sato. 2012. An inside look at google bigquery. *White paper, URL: https://cloud. google. com/files/BigQueryTechnicalWP. pdf* (2012).

[70] Quirin Scheitle, Oliver Gasser, Theodor Nolte, Johanna Amann, Lexi Brent, Georg Carle, Ralph Holz, Thomas C. Schmidt, and Matthias Wählisch. 2018. The Rise of Certificate Transparency and Its Implications on the Internet Ecosystem. In *Proceedings of the Internet Measurement Conference 2018* (Boston, MA, USA) *(IMC '18)*. Association for Computing Machinery, New York, NY, USA, 343–349. https://doi.org/10.1145/3278532.3278562

[71] Kwang-Kyu Seo. 2012. A Comparison Study between Korean Cloud Service Certification Systems and US FedRAMP. *Journal of digital convergence* 10, 11 (2012), 59–65.

[72] Christopher Soghoian and Sid Stamm. 2010. Certified lies: Detecting and defeating government interception attacks against SSL. In *Proceedings of ACM Symposium on Operating Systems Principles*. 1–18.

[73] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. 2011. On the challenges in usable security lab studies: lessons learned from replicating a study on SSL warnings. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*. 1–18.

[74] Chromium Google Open Source. 2019. EV UI Moving to Page Info. https://chromium.googlesource.com/chromium/src/+/HEAD/docs/security/ev-to-page-info.md. (Accessed on 03/04/2020).

[75] Ben Stock, Giancarlo Pellegrino, Christian Rossow, Martin Johns, and Michael Backes. 2016. Hey, you have a problem: On the feasibility of large-scale web vulnerability notification. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 1015–1032.

[76] Chee Wee Tan, Izak Benbasat, and Ronald T Cenfetelli. 2008. Building citizen trust towards e-government services: do high quality websites matter?. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE, 217–217.

[77] Laura Taylor. 2014. FedRAMP: History and future direction. *IEEE Cloud Computing* 1, 3 (2014), 10–14.

[78] United States GSA. 2020. GSA/data: Assorted data from the General Services Administration. https://github.com/GSA/data. (Accessed on 05/28/2020).

[79] United States GSA. 2020. GSA/govt-urls: This repo contains USA.gov's list of government URLs that don't end in .gov or .mil. https://github.com/GSA/govt-urls. (Accessed on 05/28/2020).

[80] Benjamin VanderSloot, Johanna Amann, Matthew Bernhard, Zakir Durumeric, Michael Bailey, and J. Alex Halderman. 2016. Towards a Complete View of the Certificate Ecosystem. In *Proceedings of the 2016 Internet Measurement Conference* (Santa Monica, California, USA) *(IMC '16)*. Association for Computing Machinery, New York, NY, USA, 543–549. https://doi.org/10.1145/2987443.2987462

A APPENDIX

A.1 Case Study 1: United States of America

The United States of America issued a memorandum M-15-13 in 2015 mandating the requirement of secure https connections across all federal websites and web services that belong to the United States government. As a result, the General Services Administration publicly provides 15 datasets at the granularity of State, Federal, Local governments in addition to the 2016 End of Presidential Term (EoT) snapshot which have mostly been archived. We create an additional dataset *full-federal-removed-diff* which contains the websites obtained from a set subtraction between the Current Federal Domains & Govt. Federal Only Domains. In Table A.1 we indicate the total number of websites reachable, the number of them which support http only, those which support https and those which support loading page content in both. We also breakdown each dataset with the corresponding certificate validity and leave out the remaining unavailable websites. We also provide a detailed breakdown of the dataset with the reasons for certificate invalidity and common errors in Table A.2.

Dataset	Total	http	http and https	https	Valid Certs	Invalid Certs
Govt. State Only Domains	827	203	106	561	406	155
Govt. Native Sovereign Only Domain	53	24	15	37	27	10
rDNS Federal Snapshot	8896	142	68	3614	3370	244
Govt. Regional Only Domains	51	18	8	32	23	9
Govt. Not used Domains	2511	845	474	1509	925	584
Govt. OCSP CRL	15	12	0	0	0	0
Govt. Quasi governmental Only Domains	64	7	4	50	36	14
End of Term 2016 Snapshot	177969	16079	9190	56531	45789	10742
Censys Federal Snapshot	47909	475	203	10415	9737	678
Other Websites	14330	157	98	3382	3096	286
Govt. Federal Only Domains	391	77	39	213	159	54
Govt. Current Federal Domains	1249	32	19	892	811	81
Govt. Local Only Domains	6228	2476	1544	4751	3613	1138
DOT .MIL (Dept. of Defense)	89	10	6	36	29	7
Govt. County Only Domains	1399	534	278	883	630	253

Table A.1: Breakdown of US GSA Datasets

In Figure A.1, we compare certificate validity to the hosting provider for the government websites in each of the individual datasets provided by the United States and considered for the case study. Our results continue to indicate that a majority of the government websites are privately hosted irrespective of the level of government *i.e.* Federal, State, or county. We additionally observe that a large number of websites are unavailable and their IP address could not be resolved.

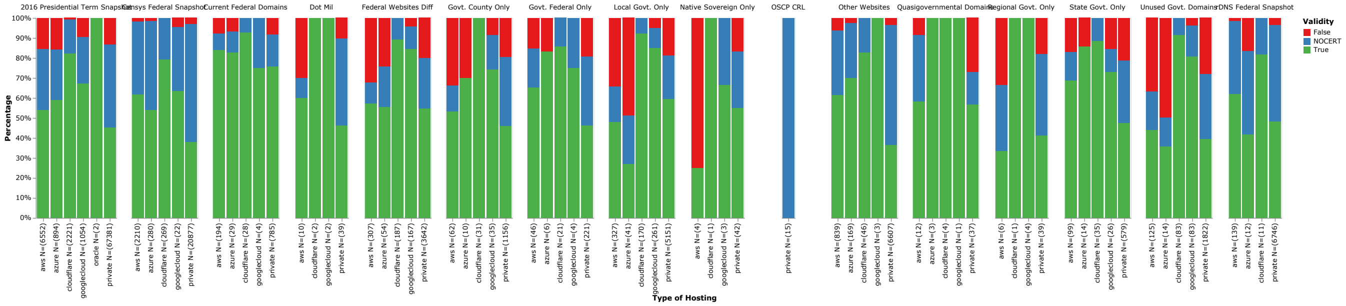

Figure A.1: Certificate Validity by Hosting Per Dataset of the GSA's Government Website Listing in the United States

In Figure A.2, we indicate the popular CAs which issue Extended Validation (EV) certificates and the number of invalid EV certificates being used by government websites in the United States. We also see all EV certs issued by Starfield Secure CA - G2, previously owned and operated by GoDaddy are invalid.

In Figure 8, we present the popular certificate issuers who issue certificates to government authorities in the United States of America. Let's Encrypt continues to be the leading certificate issuer and has much lower certificate invalidity percentage when compared worldwide.

	C **%**	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13
A	*C*	827	203	406	155	5	1	8	10	80	20	3	28	0
	P	100	24.54	49.09	18.74	0.604	0.121	0.967	1.209	9.67	2.41	0.362	3.38	0
B	*C*	53	24	27	10	0	0	1	4	5	0	0	0	0
	P	100	45.28	50.94	18.86	0	0	1.88	7.54	9.44	0	0	0	0
C	*C*	8896	142	3370	244	19	9	73	2	98	6	6	31	0
	P	100	1.59	37.88	2.74	0.2135	0.1011	0.821	0.022	1.10	0.067	0.067	0.348	0
D	*C*	51	18	23	9	0	0	1	3	4	1	0	0	0
	P	100	35.29	45.09	17.64	0	0	1.96	5.88	7.84	1.96	0	0	0
E	*C*	2511	845	925	584	16	8	27	90	249	53	19	122	0
	P	100	33.65	36.83	23.25	0.64	0.319	1.07	3.58	9.91	2.11	0.75	4.85	0
F	*C*	15	12	0	0	0	0	0	0	0	0	0	0	0
	P	100	80.0	0	0	0	0	0	0	0	0	0	0	0
G	*C*	64	7	36	14	0	0	0	0	4	6	0	4	0
	P	100	10.93	56.25	21.87	0	0	0	0	6.25	9.37	0	6.25	0
H	*C*	177969	16079	45789	10742	212	80	1320	555	5982	337	268	1419	0
	P	100	9.03	25.72	6.04	0.12	0.045	0.74	0.32	3.36	0.189	0.150	0.797	0
I	*C*	47909	475	9737	678	53	20	203	3	184	18	151	46	0
	P	100	0.99	20.32	1.41	0.11	0.04	0.42	0.006	0.384	0.037	0.315	0.096	0
J	*C*	14330	157	3096	286	15	2	44	7	173	15	15	14	1
	P	100	1.09	21.61	1.99	0.10	0.013	0.307	0.049	1.20	0.10	0.10	0.097	0.006
K	*C*	391	77	159	54	3	0	2	5	29	5	4	6	0
	P	100	19.69	40.66	13.81	0.767	0	0.511	1.27	7.41	1.27	1.02	1.53	0
L	*C*	1249	32	811	81	4	1	11	0	30	14	3	18	0
	P	100	2.56	64.93	6.48	0.32	0.08	0.88	0	2.40	1.12	0.24	1.45	0
M	*C*	6228	2476	3613	1138	34	11	89	112	584	51	34	223	0
	P	100	39.75	58.01	18.27	0.545	0.176	1.42	1.79	9.37	0.81	0.54	3.58	0
N	*C*	89	10	29	7	0	0	3	0	3	1	0	0	0
	P	100	11.23	32.58	7.86	0	0	3.37	0	3.37	1.12	0	0	0
O	*C*	1399	534	630	253	7	2	25	13	124	8	4	70	0
	P	100	38.17	45.03	18.08	0.50	0.142	1.78	0.929	8.86	0.571	0.285	5.00	0

Table A.2: Breakdown of Govt. Websites in United States by Vulnerability

- **A**: Govt. State Only Domains
- **B**: Govt. Native Sovereign Only Domains
- **C**: rDNS Federal Snapshot
- **D**: Govt. Regional Only Domains
- **E**: Govt. Not used Domains
- **F**: Govt. OCSP CRL
- **G**: Govt. Quasi governmental Only Domains
- **H**: End of Term 2016 Snapshot
- **I**: Censys Federal Snapshot
- **J**: Other Websites
- **K**: Govt. Federal Only Domains
- **L**: Govt. Current Federal Domains
- **M**: Govt. Local Only Domains
- **N**: DOT .MIL (Dept. of Defense)
- **O**: Govt. County Only Domains
- **E1**: Total Number of Domains
- **E2**: Number of HTTP Only Domains
- **E3**: Number of Valid HTTPS Domains
- **E4**: Number of Invalid HTTPS Domains
- **E5**: Certificate Has Expired Error
- **E6**: Self signed certificate in certificate chain
- **E7**: Unable to get local issuer certificate
- **E8**: Self signed certificate
- **E9**: Hostname Mismatch
- **E10**: Operation timed out
- **E11**: Connection refused
- **E12**: Unknown Exception
- **E13**: IP Address mismatch
- *C*: Count
- *P*: Percentage

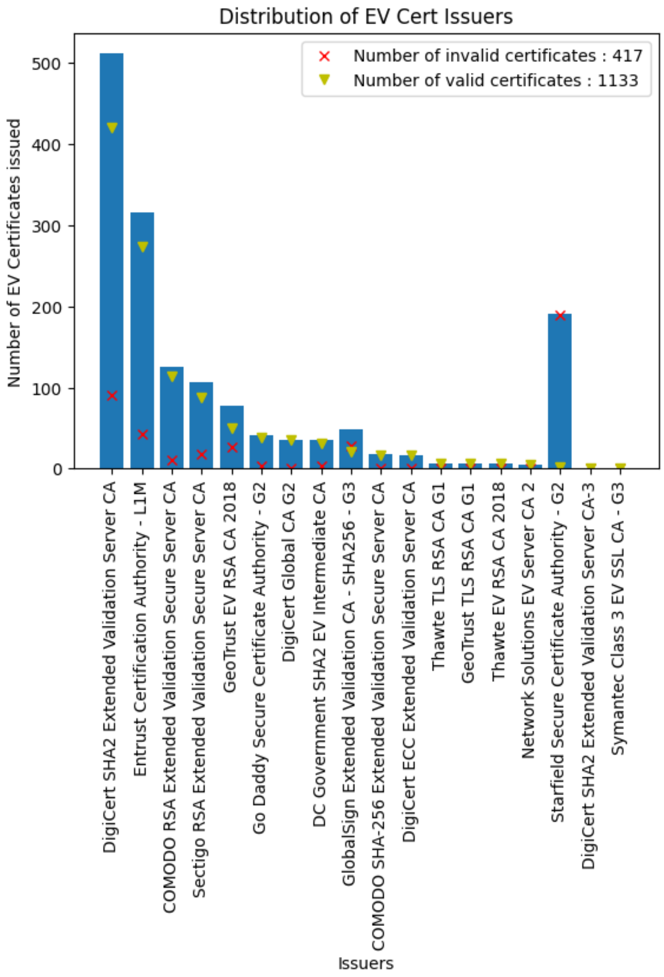

Figure A.2: Top CAs issuing EV certificates for government websites considered in the USA dataset

A.2 Case Study 2: Republic of Korea (ROK)

As shown in section 6, Republic of Korea (ROK) is very different from the United States despite having a comparably high human development index score and Internet adoption rate. In Table A.3, we present a breakdown of websites which serve content using `http`, `https` and over both. Similarly, we also provide the detailed breakdown showing the reasons for invalidity in Table A.4. Hostname mismatches continue to be the most common reasons for certificate invalidity and indicate possible misconfigurations by the system administrators/webmasters which could be easily corrected.

Dataset	Total	http	http and https	https	Valid Certs	Invalid Certs
South Korea Domains Set	21818	16814	11685	13768	5226	8542

Table A.3: Breakdown of South Korea Datasets

In Figure A.3, we show popular CAs which issue EV certificates to government websites in South Korea. The largest EV certificate provider for South Korean government websites, Thawte EV RSA CA, has ≥95% certificate invalidity. The largest number of valid EV certificates for ROK are issued by GeoTrust.

In Figure 11, we show the popular CAs used by South Korean government websites. The leading CAs for ROK are different from that of the United States or from a worldwide perspective. Many government websites in ROK continue to use CA authorities which were previously a part of the NPKI infrastructure and are now considered untrusted by many popular browsers and have been removed from the trust stores.

	Count	% of Total
Total websites considered	**21,818**	**100**
➤ Content served on HTTP	16,814	77.06
➤ Content served on HTTPS	13,768	63.10
➤ Valid HTTPS Certificates	5,226	23.95
➤ Invalid HTTPS Certificates	8,542	39.15
➤ Hostname Mismatch	2529	11.59
➤ Unable to get local issuer cert	2126	9.75
➤ Unknown Exceptions	2903	13.30
➤ Usage of self-signed cert	21	0.09
➤ Certificate Expired	23	0.10
➤ Self-signed cert in chain	818	3.75
➤ Operation Timed Out	25	0.114
➤ Connection Refused	97	0.44

Table A.4: Breakdown of the South Korean Govt. websites by vulnerability

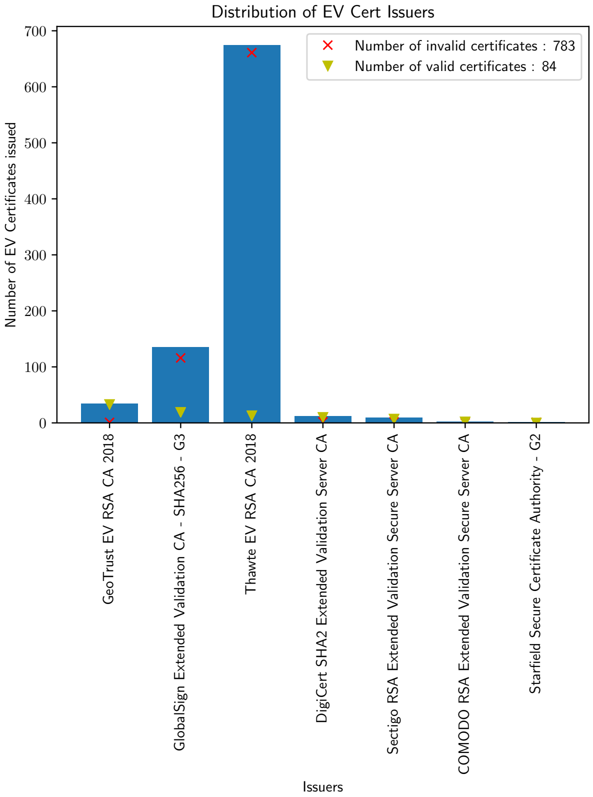

Figure A.3: Top CAs issuing EV certificates for government websites considered in the ROK dataset

A.3 Crawler

In Figure A.4, we show the effectiveness of our crawler in gathering the 134,543 unique government hostnames as described in section 4.1. The crawler is provided with the seed list of websites, visits the root page of each of the website, and uses the follow links present on the page. Only the links which have a valid country code are chosen to be further crawled and added to the queue of the crawler. We terminate the crawl after 7 levels of depth. The red line in the plot shows the percent increase in the dataset at each level, the pink line refers to the number of unique domains which are filtered at each level and compares them to the original seed list. The blue line indicates the number of domains matching the government ccTLDs.

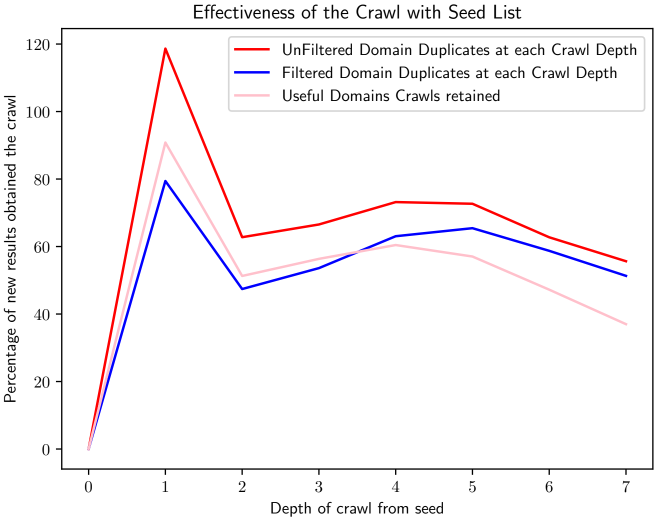

Figure A.4: Effectiveness of the Crawler in improving the seed list to gather new domains

A.4 Interconnected Governments & MITM Risks

As indicated in section 7.3, our data indicates that 75% of the countries (indicated by blue) have links from their government websites to other government websites of at least 7 countries. There are many countries (indicated in orange) being linked by government websites of ≥50 countries. A secure `https` website might provide links on their page to government websites which support only `http`, posing a risk for MITM attacks. For example, a user navigating to the `http` website from the legitimate `https` website could think that a MITM version of the `http` website served to them is legitimate and is at risk of being provided false information.

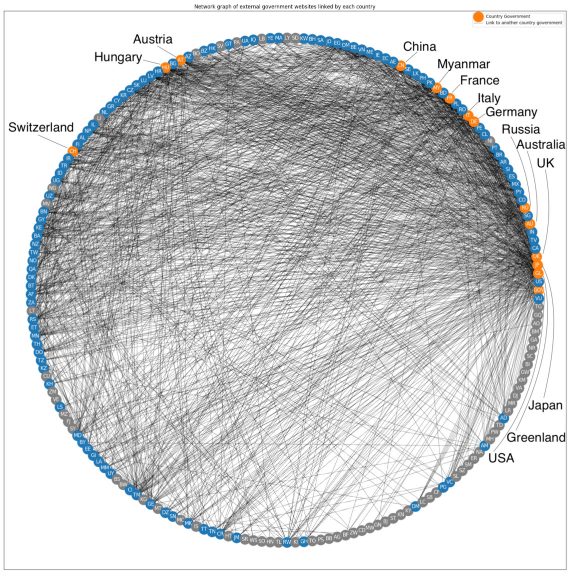

Figure A.5: Number of Government websites connected to other governments worldwide.

A.5 Worldwide EV Certificate Usage

In Figure A.6, we present popular EV certificate issuers worldwide and indicate the number of valid certificates provided by these CAs which continue to be used. As indicated in Figure 2 of the paper, we note that the most popular EV certificate issuing CA is at the 15th most popular CA. The number of invalid certificates used by the government websites, due to misconfigurations or expiries range between 15% and 20% even for EV CAs possibly indicating that the paid model of CA issuing certificates does not affect `https` validity and supports the move by popular browsers from removing explicit user interfaces which distinguished EV certificates from DV certificates [74].

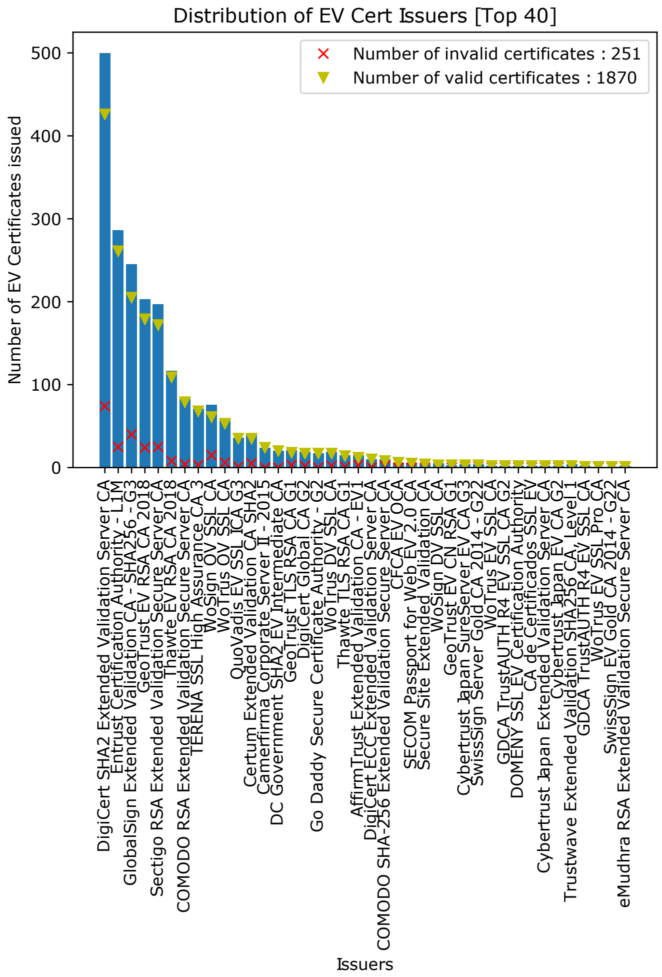

Figure A.6: Top CAs issuing EV certificates for government websites considered worldwide