# 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 old 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 certificates [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 dcb.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.

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].

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.

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

5.3.2 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, f : 285,611) with [1] 12,000 random, uniformly sampled top million non-government hostnames (mean rank: 499,206, f : 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, f : 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.

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 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].

### 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 Cloud are, 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 87.95% 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