Et tu, Brute? Privacy Analysis of Government Websites and Mobile Apps

Nayanamana Samarasinghe, Aashish Adhikari, Mohammad Mannan, Amr Youssef
Concordia University
Canada
{n_samara,a_dhika,mmannan,youssef}@ciise.concordia.ca

ABSTRACT

Past privacy measurement studies on web tracking focused on high-ranked commercial websites, as user tracking is extensively used for monetization on those sites. Conversely, governments across the globe now offer services online, which unlike commercial sites, are funded by public money, and do not generally make it to the top million website lists. As such, web tracking on those services has not been comprehensively studied, even though these services deal with privacy and security-sensitive user data, and used by a significant number of users. In this paper, we perform privacy and security measurements on government websites and Android apps: 150,244 unique websites (from 206 countries) and 1166 Android apps (from 71 countries). We found numerous commercial trackers on these services—e.g., 17% of government websites and 37% of government Android apps host Google trackers; 13% of government sites contain YouTube cookies with an expiry date in the year of 9999. 27% of government Android apps leak sensitive information (e.g., user/device identifiers, passwords, API keys) to third parties, or any network attacker (when sent over HTTP). We also found 304 government sites and 40 apps are flagged by VirusTotal as malicious. We hope our findings to help improve privacy and security of online government services, given that governments are now apparently taking Internet privacy/security seriously and imposing strict regulations on commercial sites.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**; **Software and application security**.

KEYWORDS

Government services, tracking, web, Android, privacy, security

ACM Reference Format:
Nayanamana Samarasinghe, Aashish Adhikari, Mohammad Mannan, Amr Youssef. 2022. *Et tu, Brute?* Privacy Analysis of Government Websites and Mobile Apps. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3485447.3512223

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512223

1 INTRODUCTION

Tech giants such as Google, and Facebook constantly track online user behaviors to provide a better user experience, and more importantly, to improve user profiles that they curate for monetization, e.g., via advertisements. Users are aware of, and to some extent reluctantly submit themselves to such inevitable tracking on commercial websites, to get the so-called "free" services. In contrast, one may not expect commercial trackers on government web services as those are directly funded by the tax-payers' money. Indeed, government sites are frequently used and highly trusted by users [61, 78]. Citizens use government websites and mobile apps to perform their civic obligations (e.g., taxes), query public services (e.g., waste disposal), and browse for important information (e.g., HIV, pregnancy, COVID-19). Tracking on these services could be quite revealing due to their sensitive nature [55]. If combined with user profiles on commercial sites, such tracking can make it easier to manipulate real-world user behaviors (cf. voting as targeted by Cambridge Analytica [86]). Furthermore, possible exploitation of government sites by malicious actors can also directly compromise security of users; e.g., past attacks used government sites to distribute malware, ransomware, run a botnet and cryptocurrency mining [41, 62, 74, 77, 81].

Privacy implications of web tracking have been extensively studied. Englehardt et al. [24] developed OpenWPM for large-scale evaluation of web tracking, and found Google, Facebook, Twitter and AdNexus trackers on more than 10% of Alexa top 1 million websites [3]. Privacy and security measurement studies (e.g., web tracking, HTTPS) also used sites published by Alexa, Tranco [47] and Cisco Umbrella [14], obviously due to the popularity of those top sites. However, only 9.07% government sites are present in commonly used top-million website lists [73], as government sites are used by a geographically-confined population.

Studies specifically targeting government websites (e.g., [4, 25, 79]), either focused on a specific country, or did not consider security and privacy issues in their evaluation. Recently, Singanamalla et al. [2] measured the HTTPS adoption errors and misconfigurations on government websites. In terms of tracking, a 2019 Cookiebot report [16] identified that 89% of EU government websites (e.g., nhs.uk, gov.uk) from 28 countries contained ad trackers (82% of which were from Google). However, a global perspective on commercial trackers on government services is still missing, even though governments across the world are increasingly making their services available online, especially during the current COVID-19 pandemic situation.

In this work, we perform a large-scale privacy and security measurement study on government services, using 150,244 unique government sites from 206 countries [85] and 1166 Android apps from

71 countries. We consider a website belongs to a government, if the domain name of the corresponding nameserver (ns) or mail exchanger (mx) record pertains to that government. We use a semi-automated methodology to identify 121,846 unique government domains, which we then complement with an additional 109,603 government domains from Singanamalla et al. [42, 73] (totalling 231,449 distinct domains). We then crawl the landing pages from these domains using OpenWPM [24] and measure tracking prevalence on them; a total of 150,244 domains were successfully crawled (81,205 domains were inaccessible or inactive at the time of our crawl). We leverage the content saved from government websites to identify Google Play URLs and download 1166 government Android apps (after filtering non-government apps). To understand security and privacy exposure of these apps, we use both static analysis (MobSF [53], LiteRadar [49] and Firebase scanner [68]), and dynamic analysis techniques (using a Samsung S5 phone, with Google UI/Application Exerciser Monkey). However, we limit the security evaluation of government services due to possible legal and ethical issues. In addition, we scan all government and tracking (script/cookie) domains, and government Android APKs using VirusTotal [83] to determine the possible inclusion of malicious content in government websites/apps.

We characterize widespread tracking on government services as a *betrayal* by the governments, specifically when some jurisdictions (e.g., EU, California) have explicit laws (GDPR [26], CCPA [76]) to restrict tracking on commercial sites. Similarly, the breach of trust from compromised/malicious government apps [18] is also real.

Contributions and notable findings. We develop a comprehensive framework for collecting government sites and Android apps, and a test methodology for evaluating them, primarily focusing on privacy exposure. Our main findings include:

(1) We found widespread use of commercial trackers on government websites and apps. Unsurprisingly, major trackers that are present on the regular web (cf. [71]) also dominate on government services—e.g., Google trackers are on both government sites (17%) and Android apps (37%). There were tracking cookies set to last for a long time—13% (19,566) of government sites contain YouTube cookies with an expiry date in the year of 9999. These trackers are primarily due to the inclusion of commercial content (e.g., Google maps) on government sites, and the use of analytic libraries in apps. Privacy policies of 23 (out of a selected set of 227) government sites do not mention the use of any tracker. Whether explicitly mentioned or not in policy documents, these trackers can definitely correlate user activities across commercial and government services.

(2) Government services from regions with strong privacy regulations such as the EU countries and the state of California (crawled from a localized VPN), also contain a lot of trackers, apparently violating their own regulations (GDPR and CCPA, respectively). For example, 49% (953/1942) and 69% (306/444) of EU and California government websites include known tracking scripts, respectively. These sites also include known tracking cookies with long validity periods; e.g., a total of 35 sites from both regions include known tracking cookies that are valid for 7984 years. Note that our crawler does not click on the cookie consent prompts, if present.

(3) Surprisingly, there are government services that are apparently malicious, or load content from domains labelled as malicious as per VirusTotal (see Sec. 4.3); 304 government sites and 40 governments apps are labelled as malicious. In addition, 21 tracking domains (19 included in 377 sites, 2 included in 2 apps) are labelled as malicious.

(4) Several government apps leak privacy/security sensitive information to trackers, or any network attacker. Examples: 23.1% (269/1166) of the apps expose device data (e.g., device model, device ID) to trackers; 7 apps send user login information in cleartext; 11 apps include hard-coded user/admin credentials and API keys; and 30 apps expose their unprotected Firebase datastores (apparently including confidential and personally identifiable information).

(5) Sensitive user or government data may cross jurisdictional boundaries due to the use CDNs and hosting providers. Notable examples: US/Delaware's election website elections.delaware.gov is hosted in the UK, Australia's army.defencejobs.gov.au and Somalia's centralbank.gov.so, as.parliament.gov.so are hosted in the US.

(6) We found 23 government sites from 7 countries include FullStory [1, 29] third-party script, which is used to collect the full user session (e.g., for debugging, replaying). Moreover, 5 sites expose user information (e.g., email address, search terms) to FullStory, although FullStory can be configured to limit such exposure.

We disclosed our findings on the leakage of user/admin credentials and API keys to the developers of those 11 government Android apps, but received only one response after several months (we also made several follow ups). We also reported 8 government websites flagged as malicious (by at least 5 VirusTotal engines) to site administrators/contacts of those sites, but received no response. Furthermore, we reported 38 government Android apps flagged as malicious (by at least by 1 VirusTotal engine, as the number of apps is smaller compared to government sites) to its developers, but only one developer reached out to us.

2 RELATED WORK

Tracking on popular websites. There exist a significant number of papers (e.g., [7, 19, 24, 27, 28, 50, 69]) on web tracking on popular websites. Englehardt et al. [24] developed the OpenWPM framework [63] to measure the prevalence of tracking on websites at a large scale. OpenWPM can measure both stateful (third-party scripts and cookies), and stateless (fingerprinting) tracking. Englehardt et al. found that only a few third-party tracking and advertising scripts (i.e., Google, Facebook, Twitter, Amazon, AdNexus, Oracle) were present in more than 10% of the top-1M Alexa sites. Their findings also include the use of sophisticated fingerprinting techniques (e.g., WebRTC-based, AudioContext, Battery API) in top-1M Alexa sites. The additional functionalities offered by HTML5 APIs increased the effectiveness of browser fingerprinting techniques [32]. Previous work [27, 28, 69] has also studied web tracking using popular Alexa sites from a global perspective, and found differences based on geo-location and other factors (e.g., availability of data privacy policies, laws, censorship, surveillance). Hu et al. [40] found 80% of Alexa top-2K global sites contained Google trackers. Karaj et al. [45] found third-party Google scripts in 82% of web traffic (measured using crowd-sourcing efforts). Sanchez-Rola et al. [70] observed Google tracking cookies on 93% of popular sites (on the Tranco list). We use existing methodologies and tools (e.g., OpenWPM) to specifically study commercial trackers on government sites from across the world; 91% (123,115/135,408) of these sites are not ranked in popular lists (e.g., Alexa, Cisco, Tranco).

Tracking consent solutions. Online tracking consent solutions, such as Cookiebot [16], assist website owners to manage tracking activities (i.e., detect and block trackers until a user grants consent), and ensure that web tracking complies with existing data protection regulations such as the EU GDPR. Websites integrated with Cookiebot present cookie consent banners to record user preference (accept/reject cookies). Cookiebot can also measure tracking on a given website (without an integration), and was used to analyze government websites from the 28 EU member countries; over 100 unique trackers were found. Many of these trackers were from Google (82%); only Spanish, German and Dutch government sites did not contain any tracker [8]. We found that all countries in the European Union had known tracking cookies on the analyzed government websites (291 unique trackers in total). Websites also actively take measure against users who choose not to allow cookies, e.g., by deploying aggressive browser fingerprinting techniques (see e.g., [59]). We focus on governments across the globe, and study the presence of commercial trackers on government sites, and also evaluate privacy and security issues in government Android apps.

Tracking in mobile apps. Due to the popularity of mobile apps, they also have been analyzed for privacy and security issues in the recent past, with a focus on the increasing use of tracking SDKs. Reuben et al. [9] studied 959,000 apps from US and UK Google Play stores, and found that third party tracking follows a long tail distribution dominated by Google (87.75%). Nguyen et al. [54] performed a large-scale measurement on Android apps (no mention of government apps) to understand violation of GDPR's explicit consent. The authors found 28.8% (24,838/86,163) of apps sent data to ad-related domains without explicit user consent. Several recent studies (e.g., [13, 67]) also analyzed COVID-19 tracing apps, and highlighted privacy and surveillance risks in these apps. In contrast, we target 1166 government apps of various types (including COVID-19 tracing apps) from 71 countries and territories around the globe.

HTTPS inconsistencies on government websites. There have been numerous large-scale studies on HTTPS/TLS in general. Singanamalla et al. [73] conducted the first measurement study on the HTTPS adoption in 135,408 government websites, and found a lower adoption rate (39%) compared to commercial websites; we also found similar results (61,679/150,244, 41%).[1] They also observed the prevalence of HTTPS adoption errors (e.g., the use of insecure cryptographic protocols and keys) on these sites.

Privacy and security issues on government websites. Lapses in government websites that lead into privacy and security issues have been studied for specific countries. Csontos et al. [17] found 52% of the analyzed Hungarian public sector websites used outdated server software versions and programming language releases; less than half of those websites used HTTP. The office of the auditor general in Western Australia [57] found 328 weaknesses in information technology processes (e.g., information security, IT operations, business continuity) used by 50 local government entities, out of which 10% were rated as significant. We focus on finding privacy and security issues (e.g., third party tracking, inclusion of content from malicious domains) of government services across the world.

[1]Note that as of September 2021, according to the Google Transparency Report (https://transparencyreport.google.com/https/overview?hl=en), 95% sites are now loaded over HTTPS on Chrome.

3 METHODOLOGY

In this section, we first provide details of our government website and app collection methodology. Then, we detail our privacy analysis and measurement techniques for the collected websites and Android apps; see Figure 1 for an overview of our methodology.

For websites, we define *known trackers* as the third parties (e.g., script/cookies on first-party websites) blacklisted by EasyList and EasyPrivacy [23] filtering rules; we define the rest as *unknown trackers*. We count trackers sharing the same domain name with different sub-domains separately. Furthermore, we define Android SDKs identified as trackers by MobSF [53] as *known trackers*.

3.1 Collecting government sites and apps

We compile a list of government websites from 206 countries and territories by initially using a seed list, and then refining and extending it via automated searching and crawling (between July and October, 2020). We then augment our list with the website dataset from Singanamalla et al. [42, 73]; note that our site collection methodology was developed independently.

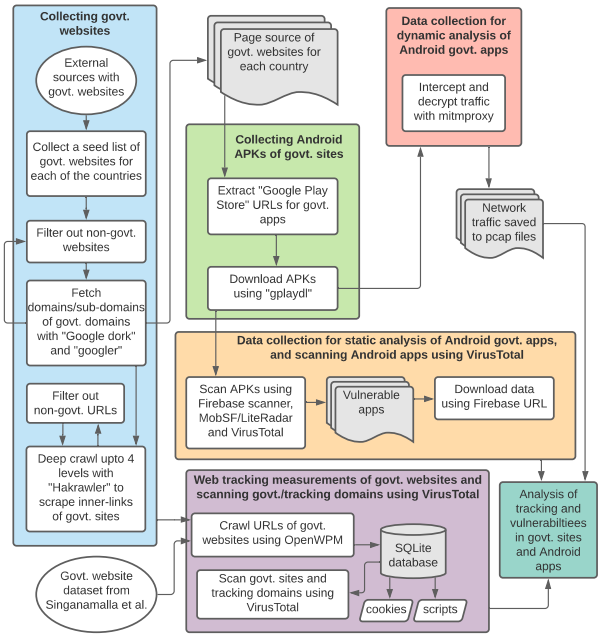


Figure 1: Overall methodology: website and Android app collection, tracking measurement on websites, and privacy and security analysis techniques used on apps.

Preparing the seed list. We begin by extracting an initial seed list of 14,861 government websites using several known sources [22, 30, 48], after removing obvious non-government entities (e.g., political parties). To eliminate any remaining non-government sites, we use `nslookup` [20] to query the nameserver (`ns`) and mail exchanger[2] (`mx`) records for each site. We then check for unique top-level domains and second/third level domains as used by various governments [84]; we then eliminate the sites that do not contain these domain suffixes in ns and mx records.

[2]Some government domains appear only in mx records.

Extending the seed list. We extract the suffixes from the seed list and prepare a Google dork [72] (e.g., site:"gov.uk") for each country. Then we use googler [88] (a command line Google search tool) to perform Google search on each Google dork and extract the search results, which may contain new domains and sub-domains. Then, we remove non-government domains from search results as explained in the previous step. We collected a total of 56,766 unique government domains/sub-domains at the end of this step.

Deep crawling to scrape inner-links. Since landing pages and inner pages of government domains collected in the previous step may contain links to other government sites, we perform a deep crawl to scrape links in the HTML page source, up to a depth of 4 levels. For this purpose, we use Hakrawler [46], that can find links in page source and the associated JavaScript files of crawled URLs. We randomize the URLs fed to Hakrawler to avoid generating a large amount of traffic to any particular web server hosting government sites. Hakrawler crawls only the web content hosted on government domains/sub-domains — i.e., it does not crawl any external websites (e.g., social media sites). For all links collected up to a depth of 4 levels, we filter out the following: links to common file extensions (e.g., docx, pdf, xls); links to social media websites; non-responsive links using *curl* [21]; domains not ending with known suffixes of government domains. After filtering, we obtained 15,214,100 URLs from 121,846 unique government domains. For each domain's landing page, we use *curl* to save the page source, which is later used to extract Google Play store URLs of government apps.

Complementing websites from Singanamalla et al. We add 135,416 government websites from Singanamalla et al. [42, 73] (collected using a different methodology including crowd sourcing via Amazon MTurk). After eliminating the overlaps, we had a total of 231,449 government websites, and we finally used 150,244 websites as the rest were unreachable (for various reasons, including unresponsive or unreachable servers). The top-10 countries with the highest number of websites in our dataset have a cumulative of 60% (90,047/150,244) — e.g., US (22,506, 15%), China (12,583, 8.4%), Bangladesh (12,258, 8.2%). We observe that 8.6% (12,873/150,244) domains of government websites make it into the Tranco [47] top-1M websites (cf. 9.07% in [73]). We manually verify our government website dataset (with a limited sample size of 100, selected randomly) to ensure false positives are eliminated. We summarize the regions and website counts in Appendix A.

Government Android apps. Government apps do not follow a common package naming convention. Therefore, we look for URLs relating to Google Play store (i.e., https://play.google.com) in the page source of government URLs saved for each country. However, not all such Google Play store URLs point to government apps (some third-party apps are also linked). We run each Google Play URL with the *curl* command to fetch developer email, developer website and privacy policy website URLs. We label a Google Play URL as a government app URL in the following cases: (i) the developer email, or developer website/privacy policy URL contains *.gov.*; (ii) the developer website/privacy policy URL appears in the list of our government websites. Then for each of the government Google Play store URLs (a total of 1566), we attempt to download the app using *gplaydl* [65]. A significant number of Android apps failed to download as they are region-locked. In the end, we collected 1166 government Android apps from 71 countries. The top-10 countries with the highest number of government Android apps in our dataset have a cumulative of 641 (out of 1166, 55%) apps — e.g., India (95, 8.1%), Australia (92, 7.9%), Indonesia (91, 7.8%).

3.2 Measurement of trackers on govt. sites

We configure the OpenWPM [24] web privacy measurement framework to launch 15 parallel browser instances in headless mode. To simulate the first visit to a website, we clear the browser profile after each URL visit. We use two Azure VMs running Ubuntu server 18.04 LTS, 4vCPUs, 16GB RAM, 30GB SSD, and a physical machine running Ubuntu 18.04 LTS, Intel Core i7-9700K, 8GB RAM, 1TB HDD for our OpenWPM measurements between Nov. 5–9, 2020. A total of 150,244 websites were successfully crawled by OpenWPM (out of 231,449), and the rest (81,205) were unreachable during our crawl (e.g., website no longer exists, SSL/TLS errors, name resolution failure, disconnection by the remote-end, timeout).

The instrumented tracking metrics from OpenWPM which include HTTP request/response of both the landing page and associated third party scripts, third party cookies, fingerprinting API calls, call stack information of web requests, and DNS resolution information are saved to a SQLite database. The saved information in OpenWPM contains both stateful (i.e., scripts and cookies) and stateless (fingerprinting) forms of metrics. We then check the saved tracking scripts and cookies for third party domains; i.e., domains of scripts/cookies that do not match the domain of the government site that they are on. We also study the known tracking scripts to find techniques used for other purposes such as session replaying and web analytics (which also could directly aid user tracking).

In order to find the correlation between privacy regulations (i.e., GDPR [26], CCPA [76]) and tracking, we separately run OpenWPM with 444 California government websites (from a VPN in California), and 1942 European Union government websites (from a VPN in the Netherlands). Note that our initial OpenWPM measurements are not done using VPNs. Our OpenWPM automation does not interact with crawled government websites, e.g., to accept or reject the cookie banners on EU sites. Therefore, our automation does not accept cookie banners on sites crawled.

3.3 Malicious govt. and tracking domains

We scan domains of all known tracking scripts/cookies in government domains (150,244), and government domains with VirusTotal to check if any of these domains are labelled as malicious. Note that, at least in some cases, VirusTotal engines[3] may misclassify or delay in updating domain categorization labels [60]. Therefore, to improve our labelling, we also automatically collect and use domain categories (e.g., phishing, malicious, spam, and advertisements, as assigned by different anti-virus engines), and community comments in VirusTotal[4] (sometimes with links to detailed analysis).

3.4 Android apps analysis

Tracking SDK detection. We use Mobile Security Framework (MobSF [53]) to find tracking SDKs embedded in government apps

[3] https://support.virustotal.com/hc/en-us/articles/115002146809-Contributors (we exclude CRDF and Quttera for their unreliable results as we observed).

[4] We used the VirusTotal API to extract community comments — see https://developers.virustotal.com/v3.0/reference#comments. To analyze the the community comments on malicious behaviour, we matched them with pre-determined keywords (e.g., phishing)

(via static analysis). We load each app to the MobSF server, scan it using the MobSF REST API, and download the JSON formatted results, which include known tracking SDKs, and strings with sensitive data and dangerous permissions [6] (e.g., camera, contacts, microphone, SMS, storage and location) used by the apps. We then use LiteRadar to find the purpose of the included tracking SDKs (e.g., Development Aid, Mobile Analytics). Finally, we store these results in a local database for our analysis.

Misconfigured Firebase database. Many Android apps, including government apps, use Google Firebase [35] (a widely used data store for mobile apps) to manage their backend infrastructure. However, due to possible misconfiguration, Android apps connected to Firebase database can be vulnerable (see e.g., [11]). Exposed data from Firebase vulnerabilities includes personally identifiable information (PII), private health information and plain text passwords [10]. Firebase scanner [68] is used to find Firebase vulnerabilities of an app (if exists). We run the Firebase scanner [68] on each APK file, which identifies the vulnerable Firebase URLs; we then download the exposed data from the Firebase datastore URL[5] and check for apparent sensitive and PII items, including: user/admin identifiers, passwords, email addresses, phone numbers. However, for ethical/legal considerations, we do not validate the leaked information (e.g., login to an app using the leaked user/admin credentials). Then we remove the downloaded Firebase datastore. We also promptly notify the developers of affected apps.

Dynamic analysis. We use a rooted Samsung S5 neo mobile phone with Android 7. We restrict only newly installed apps to proxy the traffic via mitmproxy [51] using ProxyDroid [36], to avoid collecting traffic from system and other apps. A mitmproxy root certificate is installed on the phone. We also installed mitmproxy on a separate desktop machine to collect and decrypt HTTPS traffic. Both the desktop machine and phone are connected to the same Wi-Fi network. We use adb [33] to automate the installation, launch, and uninstallation of the apps. We also use Monkey [34] with 5000 events (e.g., touch, slide, swipe, click) for each app. The network traffic is captured and stored in pcap files. We use the captured network traffic to determine sensitive information (e.g., device identifiers sent to trackers, leaked hardcoded user/admin credentials and API keys) sent to external entities. We close mitmproxy and uninstall that government app before moving to the next app.

Malicious domains and apps. We scan the APK files of 1166 government Android apps with VirusTotal. We also scan domains included in apps (as found in the network traffic) with VirusTotal.

3.5 Ethical considerations and limitations

During deep crawling to scrape inner-links to other government sites, we randomize the URLs fed to the crawler, to avoid generating a large amount of traffic to any web server hosting a government site. We do not use the sensitive information (e.g., user identifiers and passwords) extracted from static and dynamic analyses of Android apps for any intrusive validations that may have an impact to the privacy of users. In addition, we did not retain any data from exposed Firebase databases. We also reached out to the internal Research Ethics Unit of our University, and explained our experiments. They approved our methodology without requiring a full ethics evaluation. We also kept them informed about our findings and contact attempts with app developers.

Obviously, our dataset does not include all the government websites and apps available throughout the world. Furthermore, during our crawling process, we may not have encountered all trackers that are time dependent [69]. We use EasyList/EasyPrivacy [23] to filter third parties (e.g., trackers, advertisers) in government websites. Some of these filtered third parties may operate in an advertising context and may not necessarily track users, or vice-versa. It is also possible that third parties blocked by EasyList rules perform the dual role of advertising and tracking. However, the presence of third-party ad/annoyance domains is not expected on government sites as government services do not rely on ad revenue. Also government websites may intentionally use third-party scripts for tracking/analytics, and we still label such activities as tracking, as there is no technical barrier for these third-parties to use analytics data also for tracking/profiling. Determining the geolocation using IP address (see Appendix E) may not be accurate in some cases (e.g., CDN-fronted websites, non-CDN websites with multiple regional servers behind a load-balancer). However, this is less of a concern for our country-level attribution; e.g., Gharaibeh et al. [31] reported 95.8% accuracy for country-level IP-geolocation. We crawled government sites from a location outside of their home countries, except for government sites pertaining to the country where the crawler is located (i.e., Canada, the Netherlands, California). Government sites of some countries (e.g., Egypt, Iran), may not properly function when accessed from outside of the country. Also, we particularly focus on Android apps due to its larger market share, and do not consider iOS apps for this paper.[6] Android apps with obfuscated code may have impacted our static analysis, but not so on our dynamic analysis. In addition, during the dynamic analysis of apps, we did not collect traffic for those apps using SSL pinning (as we could not automatically perform un-pinning).

We involve manual steps in our methodology for verification, only when automation is unreliable or challenging (e.g., verify websites crawled pertain only to governments), to ensure that our results are reliable.

4 RESULTS: GOVERNMENT WEBSITES

In this section, we summarize our main findings on tracking and security issues on government sites. We also report additional results on use of fingerprinting APIs in Appendix F.

4.1 Third-party tracking scripts

We found 29.9% (44,880/150,244) of government websites had one or more known trackers on their landing pages, and a total of 748 unique known trackers (524,906 total known trackers). The most common known trackers were youtube.com (19,565, 13% of websites), doubleclick.net (19,339, 12.9%) and google.com (5478, 3.6%), all owned by Alphabet; see Figure 2 for the top-10 known trackers. Note that YouTube videos and Google maps are often present on government sites.

[5]The URL is of the form *<Firebase project name>.firebaseio.com/.json* (e.g., https://misenado-colombia.firebaseio.com/.json).

[6]As of August 2021, according to one estimate, Android has 72.7% market-share worldwide (https://gs.statcounter.com/os-market-share/mobile/worldwide).

We also compared the presence of third party scripts (known trackers) by country; see Figure 3a. China had a high number of government sites with known trackers (5394 sites with known trackers, out of a total of 12,583 sites, 42.9%). Russia (1623/1818, 89.3%) and Tajikistan (10/11, 90.9%) also had a high percentage of government websites with known trackers.

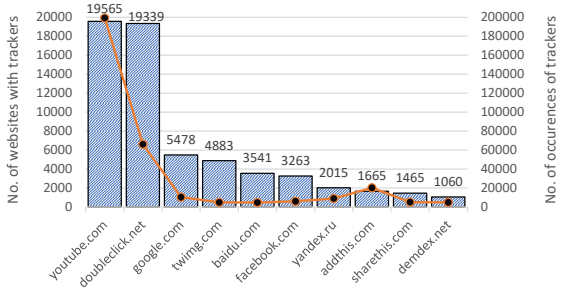


Figure 2: Top-10 known third-party tracking script sources on government sites — the bars show the number of government sites with trackers (vertical axis to the left), while the line chart shows the number of occurrences of trackers (vertical axis to the right).

We evaluated the percentages of government websites with known third-party tracking scripts for countries in different regions — see Figure 4. Notably, government websites in countries in the European Union have a relatively low percentage (14.1%) of tracking scripts, perhaps due to GDPR (although this number should be zero as per GDPR requirements). However, 49% (953/1942) European Union government websites include known tracking scripts, compared to 69% (306/444) of that of California government websites — see Appendix B. We also compared the known trackers and unknown trackers hosting tracking scripts per geographic region — see Appendix C. The proportion of known trackers is high in Africa (25,394 from a total of 27,941 trackers, 90.9%), while in South America, the proportion of unknown trackers is low (40,247/101,259, 39.7%).

Session replay by FullStory third-party script. We found some government sites in Poland (11), Mexico (1), New Zealand (1), Saudi Arabia (2), Australia (3), United States (4) and Ukraine (1) include the FullStory [1, 29] third-party script (fs.js). This script attaches event listeners to capture various events, including: button clicks, mouse movements, scrolling/resizing of windows, touch events in mobile browsers, key presses, page navigations, and network requests; all recorded events are then sent to FullStory servers. The script offers privacy options to exclude specific page elements with sensitive information (e.g., passwords, credit card numbers) to be collected/sent to FullStory servers. However, several government sites do not leverage these options, and thereby expose sensitive user information to FullStory. Examples include: my.nzte.govt.nz exposes a user's first/last name, email address during account creation; durangodigital.gob.mx exposes email address during login; several sites (e.g., rockvillemd.gov, connection.homebaseiowa.gov, nassauparadiseisland.com) expose search terms to FullStory; and several sites (e.g., eservice.sba.gov.sa, mybusiness.service.nsw.gov.au) also send browser fingerprinting information (e.g., ScreenWidth, ScreenHeight), and links clicked by users to FullStory. In contrast, parliament.vic.gov.au blocks sending search terms to FullStory.

4.2 Third-party cookies

We found many third party persistent cookies (i.e., cookies that do not expire after a session) set by known trackers, with varying validity periods; see Table 1. YouTube is the most common tracking cookie set in a large number of government sites (56,444 out of 150,244 government sites, 37.6%). About 11.5% (17,312) of government sites included cookies set by YouTube that expired within a month. YouTube cookies on 13% (19,566) of government sites are set to expire in the year 9999. Cookies set by YouTube are used to associate site visits with a Google account (if logged in) and contain information on browsing behaviours of users [87]. Also, doubleclick.net cookies on government sites (18,219, 12%) were set to expire between 1-5 years. 14 known trackers set cookies with over 5-year expiry periods; these trackers provide services in sectors including: ads/analytics (nr-data.net, cnzz.com, rezync.com, bitrix.info, 51.la), business (gemius.pl, pixlee.co), social networking (twimg.com, ok.ru), travel (sinoptik.ua), news (cctv.com) and file sharing (radikal.ru).

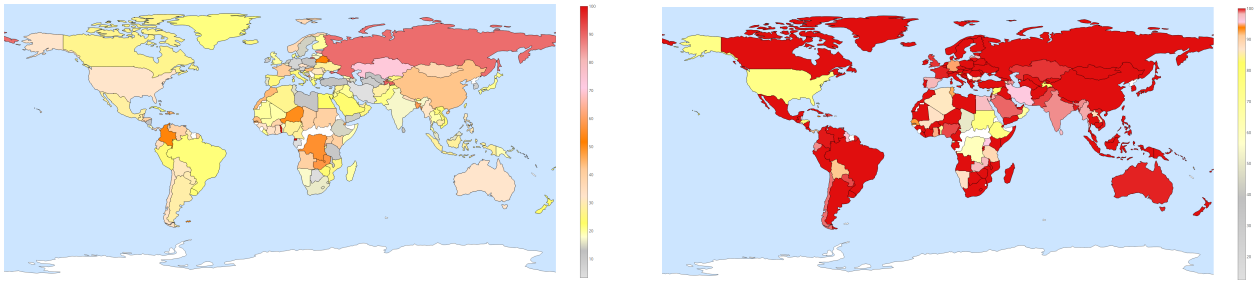
We found government websites in 112 countries set known tracking cookies on all of its websites (20,558/150,244, 13.7%). The percentage of government websites setting known tracking cookies is over 80% in 170 (out of 206) countries; see Figure 3b (also Figure 4 for region-specific prevalence of these tracking cookies). The lowest percentage of government websites with known tracking cookies was from North America (5783/7681 websites, 75.3%). The US government sites had the lowest proportion known tracking cookies (5417/7314, 74.1%), in part possibly due to California Consumer Privacy Act (CCPA) [76]. In contrast, despite GDPR [26], the percentage of government websites with known tracking cookies in the European Union was very high (2306/2355, 97.9%).

		Cookie expiry		
Tracker	**# sites**	**> 1m & ≤ 1y**	**> 1y & ≤ 5y**	**> 5y**
youtube.com	56,444	19,566	0	19,566
doubleclick.net	37,632	50	18,219	18
google.com	7731	5439	130	1
yandex.ru	4113	1995	81	2005
addthis.com	2589	921	1665	0
adsvr.org	1045	1045	0	0
rlcdn.com	793	793	0	0
bluekai.com	779	779	0	0
tapad.com	626	626	0	0
id5-sync.com	559	278	0	0

Table 1: The top-10 known tracking cookies and their expiry periods (m=month, y=year).

4.3 Government sites and tracking domains flagged as malicious

We found 0.2% (304) government sites were flagged as suspicious or malicious by VirusTotal (at least by one engine). We skipped the sites flagged as malicious by *Quttera* and *CRDF* VirusTotal engines, as the categorization returned by those engines were inconsistent. In addition, we only considered the sites that apparently were used for malicious purposes according to VirusTotal category labels and community comments, containing keywords, including: malware (41 domains), compromised (51), infection (71) spyware (36), fraud (6), weapons (3), command and control (5), bot networks (2), and



(a) third-party scripts (b) third-party cookies

Figure 3: Heatmap of percentage of government websites with known trackers in different countries (countries in white had no trackers).

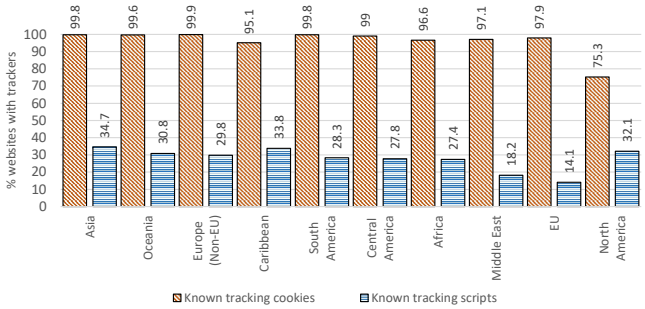


Figure 4: Known trackers (third-party scripts/cookies) on government sites by region.

callhome (4). Top 3 countries with sites flagged as malicious include Indonesia (112 out of 304), China (30) and the US (14); example sites include Royal Thai air force (rtaf.mi.th), Palestine civil defence (pcd.gov.ps), Iran health insurance organization (ihio.gov.ir) and Yemen parliament (yemenparliament.gov.ye).

We also found 15 malicious domains host known tracking scripts in 377 government sites as per VirusTotal (at least by one engine); see Table 4 (in the Appendix). We used the same procedure as for government sites to scan tracking domains with VirusTotal. 8 (out of 51) malicious domains set cookies on 311 government sites; see Table 5 (in the Appendix). We observed 50bang.org set cookies on 299 government sites.

5 RESULTS: GOVERNMENT ANDROID APPS

In this section, we present privacy and security issues found in government Android apps using static and dynamic analysis methods.

Static analysis results: Tracking SDKs and exposed databases. From MobSF, we found a total of 1647 tracking SDKs (59 unique) in 1166 apps. With LiteRadar, we checked the usage types of these SDKs (e.g., *Google Mobile Services* is used as a development aid). Similar to government websites, most tracking SDKs were also from Google (611/1647, 37.1%). Other tracking SDKs include Facebook (105/1647, 6.4%), Microsoft (34/1647, 2.1%) and One Signal (48/1647, 2.9%). Note that Google tracking SDKs are used for ad and mobile analytics. Although the collection of analytics can help provide a better user experience and improved protection (e.g., fraud detection [58]), it can also be effectively used for tracking/profiling.

We found that 2.57% (30/1166) government Android apps possibly exposed their Firebase databases with sensitive user information (as apparent from the data types); however, we did not verify/use/store this info (deleted immediately after checking the data types). Notable examples: an official app of the Colombian senate (gov.senado.app), and a real-estate regulation app from the government of Saudi Arabia (sa.housing.mullak) apparently leak user names and passwords.

Dynamic analysis results. Here we report the vulnerabilities found by inspecting the pcap files collected from our dynamic analysis (see Section 3.4). 7 apps from Bangladesh, Brazil, India, Malaysia, Nigeria, Palestine, United Arab Emirates sent login information over clear text via HTTP. These apps provide various services, including: crowd funding (com.synesis.donationapp in Bangladesh); provisioning birth/death/marriage certificates, and property tax details (in.gov.lsgkerala.mgov, in Kerala, India); services for teachers (com.trcn.teachers, in Nigeria); anti-drug volunteer management (my.gov.onegovappstore.skuadaadk, in Malaysia); and salary payments and other services for government employees (ps.gov.mtit.mservices, in Gaza, Palestine). One of these applications (com.trcn.teachers) sent traffic in the clear to an IP address belonging to an advertising/marketing service.

From the decrypted traffic from our mitmproxy, we observed 11 apps leaked hard-coded (default) user/admin credentials and API keys; see Table 2. We disclosed our findings to the app developers, and one replied mentioning that the credentials we observed were for an experimental feature which is now discontinued. For ethical/legal considerations, we did not use the leaked passwords observed for any form of validation. The services offered by these apps include crowd funding, information of leisure activities at beaches and parks, driver training and road rules, lodging of complaints, and provision of various digital resources.

We also found 23.1% (269/1166) government Android apps sent device data such as device model, and device ID to known trackers. Such device data can be used to passively track users by fingerprinting their devices. Most data types used for tracking were collected by *Branch Metrics* (device ID, device model, IPV4, screen DPI, height, and width) and *Unity Technologies* (device ID, device model, hardware name, screen DPI, height, and width).

Govt. apps and 3rd-party domains flagged as malicious. 40/1166 government apps from 22 countries were flagged as malicious by VirusTotal (at least by one engine). 10 of these apps contained a stealthy malware [39] disguising as a legitimate process executing harmful tasks (one of these apps removed the malware in a newer version), 3 apps included a stealthy adware showing as an ad blocker for Android devices (Android.WIN32.FakeAdBlocker.a), 2 apps included obfuscated malicious software that installs other

Info leak	Country	App type
Default/admin user ID and password	Australia	Parking info/directions to public hospitals
	Bangladesh	Crowd funding platform for nation building (Ek Desh)
	Brazil	Quality information of beaches
	Cambodia	Info on new driver training and road rules
	Pakistan	Communicate and provide information to public on natural disasters
	Pakistan	Lodge complaints against federal government agencies
	Portugal	The European Economic Area (EEA) program
	Singapore	A citizen-science platform for the National Parks Board (NParks)
	UK	Access services offered from local council
API key	Afghanistan	Services for urban development by Capital Region Independent Development
	Bangladesh	Used for the "Digital Bangladesh" initiative

Table 2: Exposure of sensitive information from Android apps (observed in the decrypted traffic via mitmproxy).

malware (Trojan.Trojan.Dropper.AndroidOS.Hqwar.bb). We also observed calls to 2 malicious 3rd-party domains by government apps. According to VirusTotal community comments, 2 apps (com.linkdev.dhcc.masaar and com.rajerawanna offered by United Arab Emirates and India, respectively) made calls to a malicious domain (api.ipify.org) that is infected by Cobalt Strike [52].

6 DISCUSSION

In this section, we discuss privacy implications of our findings, and list a few recommendations to mitigate these issues.

Commercial trackers. Commercial websites are heavily tracked by the top tech giants such as Google, Facebook (see e.g., [45, 70]). Both government websites and Android apps contain a significant number of such trackers; e.g., 17% of govt. sites and 37% of govt. apps contain Google trackers. Such commercial tracking is unexpected and may surprise many privacy-conscious users. Governments may want to engage citizens more actively by integrating social media resources on their websites, or attempt to understand their users' needs through the use of commercial analytical services; however, exposing their users to commercial trackers should be taken more seriously. We found 10% of the analyzed privacy policies did not even mention the use of tracking services in the corresponding government sites (see Appendix D). Government developers also need to be aware of privacy implications of using commercial JavaScript libraries and mobile SDKs, as user tracking is at the core of many of these libraries/SDKs. Clearing the browser history or the use of private browsing mode is not effective against fingerprinting attacks, which are actively being deployed to defeat cookie consent [59]. Thus, third-party scripts should be analyzed to check for the presence of any fingerprinting APIs, especially if the APIs are not essential for the service's functionality. Similarly, the use of session replay scripts (e.g., FullStory) should be avoided, or at least configured properly to reduce tracking and data exposure.

Out-sourcing app development. We found 19.8% (231/1166) apps were built by developers with non-government email addresses (137 with Gmail), indicating that at least some of these apps were developed by third-parties. Such out-sourcing may introduce the risk of leaking sensitive information, supply-chain attacks.[7]

CDNs and foreign hosting providers. Many web services, including some government services, are adopting cloud platforms (e.g., Microsoft Azure) for scalability and cost reduction. We observed several government sites that supposedly deal with sensitive user information (e.g., election, police, courts, defence, immigration, airports) were hosted in a foreign country. Privacy policies of these government sites (e.g., elections.delaware.gov–see Appendix E) do not mention anything about such outsourcing. The use of foreign hosting providers and CDNs undermine the control of the hosted data; even if the backend databases remain at a government-owned facility, user data may still be (temporarily) available to the server admins of the CDNs/hosting providers, and violate data sovereignty.[8] Although CDN hosting providers allow choosing a particular location to serve traffic, the closest location of the edge server/data center may not be within the country owning the site. There are many countries where CDNs have no data centers [12].

Malicious domains. Government sites and apps that are flagged as malicious, or include content from third parties (e.g., scripts, cookies) labelled as malicious, can harm users and diminish their trust. Unfortunately we found such malicious sites/apps on government services (304 government sites and 40 apps were flagged as malicious by VirusTotal). Governments should scan their websites/apps regularly to detect such domains.

App vulnerabilities. We found 7 government apps expose cleartext user login information, 11 apps include hard-coded (possibly admin) credentials and API keys, and 30 apps expose their unprotected Firebase datastores — all of which can enable attackers to harvest PII at a large scale. We strongly recommend developers to use HTTPS properly (cf. [73]), not to rely on cloud-hosted mobile backends such as Google Firebase (exposing user data to commercial operators), and not to include admin API keys/credentials in the app code (possibly exposing user data to anyone). Security issues regarding the use of cloud-based mobile backends have been analyzed in recent work [5, 89], and developers should check their apps and servers for similar issues.

7 CONCLUSION

Despite being publicly funded by tax payers money, government services enable commercial trackers to collect data about citizens virtually everywhere across the globe. From our analysis of 150,244 government websites and 1166 government Android apps, we found Google dominates in tracking, closely resembling the same trend as in the commercial domain, which is largely powered and monetized by tracking/profiling; cf. [40]. A downside compared to commercial services is that users have no choice in terms of switching to another provider. Finally, since many governments continue to move to digital platforms, the relevant government authorities responsible to ensure privacy in each country/region should periodically review government websites and mobile apps for tracking, privacy and security exposures, at least to comply with their own legislation.

[7] Cf. the recent SolarWinds incident: https://www.cisecurity.org/solarwinds/

[8] Several governments are considering legislation on these issues–wikipedia.org/wiki/Data_sovereignty; see also the French govt. agreement with Google and Microsoft [66].

REFERENCES

[1] Gunes Acar, Steven Englehardt, and Arvind Narayanan. 2020. No boundaries: data exfiltration by third parties embedded on web pages. *PoPETs* 2020, 4 (2020), 220–238.

[2] Eman Salem Alashwali, Pawel Szalachowski, and Andrew Martin. 2020. Exploring HTTPS Security Inconsistencies: A Cross-Regional Perspective. *Computers & Security* 97 (2020). Article number 101975.

[3] Alexa.com. 2021. Alexa Top Sites. https://aws.amazon.com/alexa-top-sites/.

[4] Abdullah Ahmed Ali and Mohd Zamri Murah. 2018. Security Assessment of Libyan Government Websites. In *Cyber Resilience Conference (CRC'18)*. Putrajaya, Malaysia.

[5] Omar Alrawi, Chaoshun Zuo, Ruian Duan, Ranjita Pai Kasturi, Zhiqiang Lin, and Brendan Saltaformaggio. 2019. The Betrayal At Cloud City: An Empirical Analysis Of Cloud-Based Mobile Backends. In *USENIX Security Symposium'19*. Santa Clara, CA, USA.

[6] Android. 2021. Permissions on Android. Online article (2021). https://developer.android.com/guide/topics/permissions/overview.

[7] Muhammad Ahmad Bashir, Sajjad Arshad, William K. Robertson, and Christo Wilson. 2016. Tracing Information flows between Ad exchanges using retargeted Ads. In *USENIX Security Symposium'16*. Austin, TX, USA.

[8] BBC. 2019. Tracking tools found on EU government and health websites. Online article (2019). https://www.bbc.com/news/technology-47624206.

[9] Reuben Binns, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert, and Nigel Shadbolt. 2018. Third party tracking in the mobile ecosystem. In *ACM WebSci'18*. Amsterdam, Netherlands.

[10] businesswire. 2018. 62% of enterprises exposed to sensitive data loss via Firebase vulnerability. Online article (2018). https://www.businesswire.com/news/home/20180619005540/en/62-of-Enterprises-Exposed-to-Sensitive-Data-Loss-via-Firebase-Vulnerability.

[11] BusinessWire.com. 2018. 62% of Enterprises Exposed to Sensitive Data Loss via Firebase Vulnerability. News article (June 19, 2018). https://www.businesswire.com/news/home/20180619005540/en/62-of-Enterprises-Exposed-to-Sensitive-Data-Loss-via-Firebase-Vulnerability.

[12] CDN Planet. 2021. Content Delivery Networks per country. Online article (2021). https://www.cdnplanet.com/geo/.

[13] Hyunghoon Cho, Daphne Ippolito, and Yun William Yu. 2020. Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. *arXiv preprint arXiv:2003.11511* (2020).

[14] Cisco. 2020. Cisco Umbrella 1 Million. Online article (2020). https://umbrella.cisco.com/blog/cisco-umbrella-1-million.

[15] Clym. 2021. How The CCPA Affects The Cookie Policy. Online article (2021). https://www.clym.io/how-the-ccpa-affects-the-cookie-policy/.

[16] Cookiebot. 2019. Ad tech surveillance on the public sector web. Online article (2019). https://www.cookiebot.com/media/1121/cookiebot-report-2019-medium-size.pdf.

[17] Balázs Csontos and István Heckl. 2021. Accessibility, usability, and security evaluation of Hungarian government websites. *Universal Access in the Information Society* 20, 1 (2021), 139–156.

[18] Cyble. 2021. Android Trojan Malware Disguised As Syrian E-Gov Android App. Online article (2021). https://blog.cyble.com/2021/05/27/android-trojan-malware-disguised-as-syrian-e-gov-android-app/.

[19] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *NDSS'19*. San Diego, CA, USA.

[20] die.net. 2010. nslookup. Online article (2010). https://linux.die.net/man/1/nslookup.

[21] die.net. 2021. curl. Online article (2021). https://linux.die.net/man/1/curl.

[22] Digital.gov. 2021. GSA govt-urls. https://github.com/GSA/govt-urls.

[23] EasyList. 2020. EasyList. Online article (2020). https://easylist.to/.

[24] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *CCS'16*. Vienna, Austria.

[25] Kristin R Eschenfelder, John C Beachboard, Charles R McClure, and Steven K Wyman. 1997. Assessing US federal government websites. *Government Information Quarterly* 14, 2 (1997), 173–189.

[26] Europa.eu. 2016. EU GDPR. Online article (2016). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN.

[27] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. 2014. The Rise of Panopticons: Examining Region-Specific Third-Party Web Tracking. In *Traffic Monitoring and Analysis (TMA'14)*. London UK.

[28] Nathaniel Fruchter, Hsin Miao, Scott Stevenson, and Rebecca Balebako. 2015. Variations in tracking in relation to geographic location. In *Web 2.0 Security and Privacy (W2SP'15)*. San Jose, CA, USA.

[29] FullStory. 2021. How does FullStory recording work to recreate my users' experience? Online article (2021). https://help.fullstory.com/hc/en-us/articles/360032975773-How-does-FullStory-recording-work-to-recreate-my-users-experience-.

[30] G. Anzinger. 2002. Worldwide Governments on the WWW. http://www.gksoft.com/govt/en/world.html.

[31] Manaf Gharaibeh, Anant Shah, Bradley Huffaker, Han Zhang, Roya Ensafi, and Christos Papadopoulos. 2017. A look at router geolocation in public and commercial databases. In *ACM Internet measurement conference (IMC'17)*. London, United Kingdom.

[32] Alejandro Gómez-Boix, Pierre Laperdrix, and Benoit Baudry. 2018. Hiding in the crowd: An analysis of the effectiveness of browser fingerprinting at large scale. In *WWW'18*. Lyon, France.

[33] Google. 2020. Android Debug Bridge (adb). Online article (2020). https://developer.android.com/studio/command-line/adb.

[34] Google. 2020. monkeyrunner. Online article (2020). https://developer.android.com/studio/test/monkeyrunner.

[35] Google. 2021. Firebase. Online article (2021). https://firebase.google.com/.

[36] Google Play. 2021. ProxyDroid. Online article (2021). https://play.google.com/store/apps/details?id=org.proxydroid&hl=en_CA&gl=US.

[37] Government of Canada. 2020. Bill C-11: An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make related and consequential amendments to other Acts. Proposed legislation: 2020; https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c11.html.

[38] Government of Canada. 2020. Personal Information Protection and Electronic Documents Act. Enacted: 2000, last amended: 2019; https://laws-lois.justice.gc.ca/ENG/ACTS/P-8.6/index.html.

[39] howtoremove.guide. 2020. Trojan.Malware.300983.susgen. Online article (2020). https://howtoremove.guide/trojan-malware-300983-susgen/.

[40] Xuehui Hu, Guillermo Suarez de Tangil, and Nishanth Sastry. 2020. Multi-country Study of Third Party Trackers from Real Browser Histories. In *IEEE EuroS&P'20*. Online.

[41] India Times. 2018. Hackers mined a fortune from Indian websites. Online article (2018). https://economictimes.indiatimes.com/small-biz/startups/newsbuzz/hackers-mined-a-fortune-from-indian-websites/articleshow/65836088.cms.

[42] Information and Communication Technology for Development (ICTD) Lab. 2020. HTTPS Adoption Measurement in Governments Worldwide. Online article (2020). https://github.com/uw-ictd/GovHTTPS-Data.

[43] Internet Society. 2018. Personal Data Protection Guidelines for Africa. Online article (2018). https://www.internetsociety.org/resources/doc/2018/personal-data-protection-guidelines-for-africa/.

[44] Internet world stats. 2021. World country list. Online article (2021). https://www.internetworldstats.com/list1.htm.

[45] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M Pujol. 2018. WhoTracks.Me: Shedding light on the opaque world of online tracking. *arXiv preprint arXiv:1804.08959* (2018).

[46] L. Stephens. 2020. Hakrawler. Online article (2020). https://github.com/hakluke/hakrawler.

[47] 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 *NDSS'19*. San Diego, CA, USA.

[48] M. Richt. 2020. German Government Domains. Online article (2020). https://github.com/robbi5/german-gov-domains/.

[49] M. Zi'ang. 2020. LiteRadar. Online article (2020). https://github.com/pkumza/LiteRadar.

[50] Max Maass, Pascal Wichmann, Henning Pridöhl, and Dominik Herrmann. 2017. Privacyscore: Improving privacy and security via crowd-sourced benchmarks of websites. In *Annual Privacy Forum (APF'17)*. Vienna, Austria.

[51] mitmproxy. 2021. mitmproxy. Online article (2021). https://mitmproxy.org/.

[52] mitre. 2021. Cobalt strike. Online article (2021). https://attack.mitre.org/software/S0154/.

[53] MobSF. 2020. Mobile Security Framework (MobSF). Online article (2020). https://github.com/MobSF/Mobile-Security-Framework-MobSF.

[54] Trung Tin Nguyen, Michael Backes, Ninja Marnau, and Ben Stock. 2021. Share First, Ask Later (or Never?)-Studying Violations of GDPR's Explicit Consent in Android Apps. In *USENIX Security Symposium (USENIX Security'21)*. Online.

[55] Joshua D Niforatos, Alexander R Zheutlin, and Jeremy B Sussman. 2021. Prevalence of Third-Party Data Tracking by US Hospital Websites. *JAMA Network Open* 4, 9 (2021), e2126121–e2126121.

[56] OECD.org. 2011. Classification of the Functions of Government (COFOG). Online article (2011). https://www.oecd.org/gov/48250728.pdf.

[57] Office of the auditor general western Australia. 2021. Local Government General Computer Controls. Online article (2021). https://audit.wa.gov.au/wp-content/uploads/2021/05/Report-23_Local-Government-General-Computer-Controls.pdf.

[58] OneSpan. 2021. Fraud Analytics. Online article (2021). https://www.onespan.com/topics/fraud-analytics.

[59] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. 2021. User Tracking in the Post-cookie Era: How Websites Bypass GDPR Consent to Track Users. In *WWW'21*. Ljubljana, Slovenia.

[60] Peng Peng, Limin Yang, Linhai Song, and Gang Wang. 2019. Opening the Blackbox of VirusTotal: Analyzing Online Phishing Scan Engines. In *IMC'19*. Amsterdam, Netherlands.

[61] Pew Research center. 2002. The rise of the e-citizen: How people use government agencies' web sites. Online article (2002). https://www.pewresearch.org/internet/2002/04/03/the-rise-of-the-e-citizen-how-people-use-government-agencies-web-sites/.

[62] Pierluigi Paganini. 2017. US Government website was hosting a JavaScript downloader delivering Cerber ransomware. Online article (2017). https://securityaffairs.co/wordpress/62629/hacking/us-government-website-malware.html.

[63] Princeton University. 2020. OpenWPM. Online article (2020). https://github.com/citp/OpenWPM.

[64] Gaston Pugliese, Christian Riess, Freya Gassmann, and Zinaida Benenson. 2020. Long-Term Observation on Browser Fingerprinting: Users' Trackability and Perspective. *PoPETs* 2020, 2 (2020), 558–577.

[65] R. Alam. 2020. gplaydl. Online article (2020). https://github.com/rehmatworks/gplaydl.

[66] Reuters. 2021. France embraces Google, Microsoft in quest to safeguard sensitive data. Online article (2021). https://www.reuters.com/technology/france-embraces-google-microsoft-quest-safeguard-sensitive-data-2021-05-17/.

[67] Frantz Rowe. 2020. Contact tracing apps and values dilemmas: A privacy paradox in a neo-liberal world. *International Journal of Information Management* 55 (2020), 102178.

[68] S. Sahni. 2019. Firebase scanner. Online article (2019). https://github.com/shivsahni/FireBaseScanner.

[69] Nayanamana Samarasinghe and Mohammad Mannan. 2019. Towards a global perspective on web tracking. *Computers & Security* 87 (2019). Article number 101569.

[70] Iskander Sanchez-Rola, Matteo Dell'Amico, Davide Balzarotti, Pierre-Antoine Vervier, and Leyla Bilge. 2021. Journey to the Center of the Cookie Ecosystem: Unraveling Actors' Roles and Relationships. In *IEEE Symposium on Security and Privacy (SP'21)*. Online.

[71] Iskander Sanchez-Rola and Igor Santos. 2018. Knockin'on trackers' door: Large-scale automatic analysis of web tracking. In *18*. Saclay, France.

[72] SecureIca. 2020. Exploring Google Hacking Techniques using Dork. Online article (2020). https://medium.com/nassec-cybersecurity-writeups/exploring-google-hacking-techniques-using-google-dork-6df5d79796cf.

[73] 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)*. Online.

[74] Softpedia news. 2013. Hacked Turkish Government website used to distribute malware. Online article (2013). https://news.softpedia.com/news/Hacked-Turkish-Government-Website-Used-to-Distribute-Malware-389937.shtml.

[75] Konstantinos Solomos, John Kristoff, Chris Kanich, and Jason Polakis. 2021. Tales of Favicons and Caches: Persistent Tracking in Modern Browsers. In *NDSS'21*. Online.

[76] State of California Department of Justice. 2021. California Consumer Privacy Act (CCPA). Online article (2021). https://oag.ca.gov/privacy/ccpa.

[77] The Guardian. 2018. Government websites hit by cryptocurrency mining malware. Online article (2018). https://www.theguardian.com/technology/2018/feb/11/government-websites-hit-by-cryptocurrency-mining-malware.

[78] Caroline J Tolbert and Karen Mossberger. 2006. The effects of e-government on trust and confidence in government. *Public administration review* 66, 3 (2006), 354–369.

[79] Andrew Tolley and Darren Mundy. 2009. Towards workable privacy for UK e-government on the web. *International Journal of Electronic Governance* 2, 1 (2009), 74–88.

[80] weareprivacy.com. 2021. Policy Highlights. Online article (2021). https://github.com/weareprivacy/policy-highlights.

[81] Vice.com. 2020. Hackers turned Virginia government websites into elaborate eBooks scam pages. Online article (2020). https://www.vice.com/en/article/88947x/hackers-virginia-government-websites-ebooks-scam.

[82] Virginia.gov. 2021. SB 1392 Consumer Data Protection Act; establishes a framework for controlling and processing personal data. Online article (2021). https://lis.virginia.gov/cgi-bin/legp604.exe?211+sum+SB1392.

[83] VirusTotal. 2021. VirusTotal. Online article (2021). https://www.virustotal.com.

[84] Wikipedia. 2021. .gov. Online article (2021). https://en.wikipedia.org/wiki/.gov.

[85] Wikipedia. 2021. List of sovereign states. Online article (2021). https://en.wikipedia.org/wiki/List_of_sovereign_states.

[86] Wired. 2019. How Cambridge Analytica Sparked the Great Privacy Awakening. Online article (2019). https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/.

[87] World mail & express americas conference. 2021. Cookie policy. Online article (2021). https://www.wmxamericas.com/cookie-policy/.

[88] Z. Wang. 2020. googler. Online article (2020). https://github.com/jarun/googler.

[89] Chaoshun Zuo, Zhiqiang Lin, and Yinqian Zhang. 2019. Why Does Your Data Leak? Uncovering the Data Leakage in Cloud from Mobile Apps. In *IEEE Symposium on Security and Privacy (SP'19)*. San Fransisco, CA, USA.

A REGIONS AND GOVERNMENT SITE COUNTS

We list the regions, and the count of government websites (in countries/territories of the corresponding regions) used in our study in Table 3.

Region	# websites
Africa	4586
Asia	60,357
Central America	2506
Europe (Non-EU)	15,148
European Union	16,681
Middle East	3209
North America	23,934
Oceania	6588
South America	15,939
Caribbean	1296

Table 3: List of regions and government website counts (countries are grouped in regions based on the categorization in [44]).

B TRACKERS ON EU AND CALIFORNIA GOVERNMENT SITES

As more services are going digital, and many commercial entities' sole business model is based on profiling users, at least some governments are apparently starting to take user privacy more seriously. They are also enacting regulations to impose significant penalties to commercial online service providers for the violation of data privacy and security measures, which include: unnecessary data collection, tracking without consent, and failing to protect personal data. Prominent regulations include: the EU General Data Protection Regulation (GDPR) [26], California Consumer Privacy Act (CCPA) [76], Virginia Consumer Data Protection Act (CDPA) [82], Personal Data Protection Guidelines for Africa [43], Canadian Personal Information Protection and Electronic Documents Act (PIPEDA) [38] (and the newly proposed legislation [37]). Ironically, many governments fail to lead by example as apparent from our results. In this section our emphasis is on the impact of GDPR/CCPA on tracking.

European Union. All websites must comply with GDPR [26] when accessed from any EU member state. GDPR is an opt-in privacy regulation (e.g., user consent must be obtained before tracking them). We found 49% (953/1942) EU government websites include known tracking scripts; note that we visit these sites via OpenWPM from a VPN in the Netherlands. Most tracking scripts (524, 27%) on these sites are served by Google, followed by Facebook (54, 2.8%), Cloudflare (24, 1.2%), and Twitter (23, 1.2%). We also observed companies (e.g., CookieLaw and Cookiebot) that provide solutions (e.g., provision of cookie banners) to adhere to GDPR, included scripts on EU government websites that are categorized as trackers by EasyList/EasyPrivacy [23]. Notably, 24 (out of 1942) government sites (e.g., Germany, Lithuania, Denmark) include tracking cookies that are valid for 7984 years; see Table 6.

California websites. Websites accessed from California are subjected to CCPA [15, 76], which is an opt-out privacy regulation. For example, CCPA does not require websites accessed from the state of California to provide explicit cookie consent (unlike GDPR). We observed 306/444 (69%) California government websites include known tracking scripts, mostly from Google (163/444), followed by CivicPlus and Microsoft (each 22, 5%), Siteimprove (13, 2.9%), and Facebook (11, 2.5%). Note that we crawled these sites from a VPN located in California. We also found website design companies serving governments (e.g., CivicPlus, Revize) included tracking scripts in government websites. In addition, 11 (2.5%) California government sites set cookies that are valid for 7984 years; see Table 7.

C THIRD PARTY SCRIPTS ON GOVERNMENT SITES BY REGION

We present the proportion of known/unknown trackers for different regions in Figure 5.

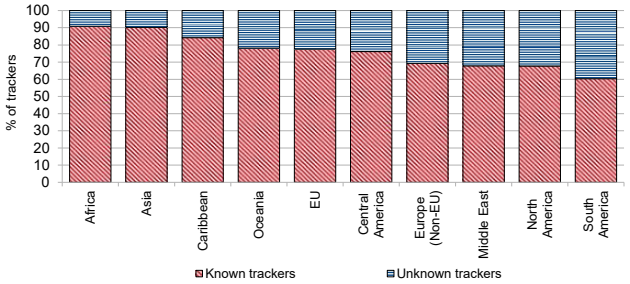


Figure 5: Proportions of third-party scripts (known trackers vs. unknown trackers) on government sites per region.

D PRIVACY POLICIES IN GOVERNMENT WEBSITES WITH TRACKERS

For this analysis we leveraged 551 privacy policy URLs extracted from the government Android apps (see Section 3.1). We found only 41.2% (227/551) of the corresponding government sites included trackers (scripts/cookies). 23/227 sites do not mention the use of tracking services in their privacy policies — based on matching the policy content with a set of predefined keywords (e.g., analytics, 3rd party, Google, Facebook, Twitter, Linkedin) using *Policy Highlights* [80]. Government sites with top unique tracking domains, but not emphasizing the use of tracking services in their privacy policies include privacy.gov.ph (8) — national privacy commission of Philippines, fsq.moh.gov.my (5) — food safety and quality division of Malaysia. The unique tracking domains in these government sites include facebook.com, facebook.net, google.com, google-analytics.com, googletagmanager.com, gstatic.com, youtube.com, ytimg.com, wp.com. There were 9.3% (21/227) privacy policies of government sites that are not written in English (we could not translate 6 of them). There were also 11 privacy policy URLs of government sites that no longer exist.

E FOREIGN-HOSTED GOVERNMENT SITES

We extract the DNS resolution information of the crawled government sites from OpenWPM to find the IP of each domain. Then using *geoiplookup*,[9] we determine the geolocation and Autonomous System Number (ASN) details of each IP address. Singanamalla et al. [73] found 94.5% (127,327/134,685) of government sites are either hosted privately or by an unknown hosting provider. In contrast, our analysis focused on government sites hosted in foreign countries.

[9] https://linux.die.net/man/1/geoiplookup

Malicious type	Tracking domains	# Govt. sites (example countries)
Malware, malnets, malvertising	iclickcdn.com, qdatasales.com, graizoah.com, 50bang.org, popcash.net	320 (China, India, Pakistan)
Browser hijacking	otrwaram.com	1 (Brazil)
Adware, unwanted redirects	newrrb.bid, supercounters.com, tradeadexchange.com	43 (Indonesia, Myanmar, Vietnam)
Potentially unwanted program	coinpot.co	4 (Bangladesh, Kyrgyzstan)
Spam	freecounter.ovh	3 (Colombia, Malaysia, Pakistan)
Suspicious	ufpcdn.com, dprtb.com, loulouly.net, adhitzads.com	6 (Indonesia, Malta, USA)

Table 4: Tracking scripts included from potentially malicious domains.

Malicious type	Tracking domains	# Govt. sites (example countries)
Malware and malnets	pingclock.net, qdatasales.com, 50bang.org	303 (China, Malaysia)
Potentially unwanted programs	iyfsearch.com, coinpot.co, rtmark.net	4 (Bangladesh, Kyrgyzstan)
Suspicious	ufpcdn.com, remarketingpixel.com	4 (Kenya)

Table 5: Tracking cookies set by potentially malicious domains.

Validity period	# sites	Example trackers
7984 years	24	iteimproveanalytics.io, snoobi.com, nr-data.net
16 years	1	trafic.ro
1 – 5 years	27	statcounter.com, omtrdc.net, adverticum.net
3 – 6 months	11	pubmatic.com, innovid.com

Table 6: Cookie validity periods on EU govt. sites.

Validity period	# sites	Example trackers
7984 years	11	siteimproveanalytics.io, rfihub.com, nr-data.net
10 years	1	webtrendslive.com
1–2 years	15	stackadapt.com, scanscout.com, rubiconproject.com
1–6 months	7	krxd.net, demdex.net

Table 7: Cookie validity periods on California govt. sites.

We observed 194 countries host their site content using services from a foreign country; e.g., 2.2% (489/22,506) websites from the United States and 2.9% (370/12,583) websites from China are hosted outside these countries. These sites are hosted by cloud providers (i.e., hosting/CDN providers) with data centers around the globe; Wix (102) and Akamai (67) host most of these sites for the United States, while Quantil (202), Cloudflare (39) and Alibaba (25) hosted most sites for China. Some countries in Africa host all their government sites (in our dataset) outside: Chad (5), Congo (9), Equatorial Guinea (2), Somalia (16), Togo (3). Most prominent government sites (10) in Somalia (e.g., centralbank.gov.so, as.parliament.gov.so) were hosted by a provider (Unitedlayer) in the US.

We analyzed 1466 government websites, which are likely to be hosted at a foreign provider, not at CDNs due to the fact that ASN names of these websites did not contain a CDN listed in [12], and their IP addresses remained static and at a foreign geolocation when accessed both from IP addresses in Canada and in the Netherlands. We also found the categories [56] of these websites by parsing the text within meta tags of request headers–to determine if these sites serve any sensitive/critical purposes. Notable categories of these sites include: election (e.g., US/Delaware's election website elections.delaware.gov hosted in the UK); defence (e.g., Australia's army.defencejobs.gov.au hosted in the US); police (e.g., Australia/Victoria's policecareer.vic.gov.au hosted in the US); courts (e.g., a New Zealand district court website: districtcourts.govt.nz hosted in the US); immigration (e.g., Papua New Guinea's immigration.gov.pg hosted in Australia); and airports (e.g., Kenya's kaa.go.ke hosted in the Netherlands).

F FINGERPRINTING APIS

We found many instances of calls to various fingerprinitng APIs on government websites. Examples include: Storage (5,355,626), CanvasRendering2D (7,615,438), window.navigator (3,349,296), HTMLCanvasElement (1,102,482), hardware related APIs[10] (230,426), window.screen (99,504), audio related APIs (16,274), window.navigator.geolocation (8334), RTC (2655). APIs related to Audio, hardware, RTC and window.screen can track users for a relatively longer period as the characteristics of those fingerprints generally remain static for a long time [64, 75]. We found other privacy implications from the fingerprinting APIs: Window.navigator.geolocation gives a website access to the location of user device (called 8334 times), and RTC is used to discover local IPs without user permission [24] (called 498 times). Such a combination of multiple fingerprinting APIs can be used to identify a user with a high precision [24], and reportedly being used to bypass EU GDPR cookie restrictions [59].

G POTENTIALLY MALICIOUS DOMAINS HOSTED BY TRACKERS

We list potentially malicious domains including scripts and cookies on government sites in Table 4 and Table 5 respectively.

[10]Hardware fingerprinting APIs include: window.navigator.hardwareConcurrency, window.navigator.mediaDevices, window.navigator.getGamepads, window.navigator.oscpu, window.navigator.platform, window.navigator.vibrate and window.navigator.maxTouchPoints.