Measuring Web Cookies in Governmental Websites

Matthias Götze
TU Berlin

Srdjan Matic
IMDEA Software Institute

Costas Iordanou
Cyprus University of Technology

Georgios Smaragdakis
TU Delft

Nikolaos Laoutaris
IMDEA Networks Institute

ABSTRACT

In recent years, governments worldwide have moved their services online to better serve their citizens. Benefits aside, this choice increases the danger of tracking via such sites. This is of great concern as governmental websites increasingly become the only interaction point with the government. In this paper, we investigate popular governmental websites across different countries and assess to what extent the visits to these sites are tracked by third-parties. Our results show that, unfortunately, tracking is a serious concern, as in some countries up to 90% of these websites create cookies of third-party trackers without any consent from users. Non-session cookies, that are created by trackers and can last for days or months, are widely present even in countries with strict user privacy laws. We also show that the above is a problem for official websites of international organizations and popular websites that inform the public about the COVID-19 pandemic.

CCS CONCEPTS

• **Information systems → World Wide Web**; • **Security and privacy → Human and societal aspects of security and privacy**.

KEYWORDS

Official Web Services; Web Cookies; User Tracking; COVID-19; GDPR.

ACM Reference Format:
Matthias Götze, Srdjan Matic, Costas Iordanou, Georgios Smaragdakis, and Nikolaos Laoutaris. 2022. Measuring Web Cookies in Governmental Websites. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3501247.3531545

1 INTRODUCTION

Electronic governance, also known as e-governance, refers to the ongoing efforts by governments around the globe to deliver government services, such as announcements, communication, exchange of information, and point of service to their citizens. Studies have shown that citizens' and companies' productivity increases when electronic governance covers a large spectrum of public services and can be accessible by a large portion of the population [29]. E-governance also has the potential to reduce the cost of a government, eliminate corruption, and drastically reduce the service time for citizens [32]. For this reason, many countries devote substantial budgets to run and enhance these services that, combined with investments in broadband access, have the potential to eliminate "digital divisions" and make government services accessible even to lower income citizens [33]. In some developed countries, several services are predominantly offered online, e.g., tax declarations or access to legal documents, and face-to-face interaction with public servants is arranged only in very exceptional cases. The investments to e-governance has proved to be extremely valuable during the COVID-19 pandemic, allowing a large fraction of the interactions between citizens and the authorities to remain uninterrupted during these difficult times [17, 34].

A potential risk from e-governance is that since it represents a unique point of interaction for mandatory and indispensable services for all citizens, it can, unintentionally or not, become a single point of monitoring and tracking for the entire population of a country. A readily available way to achieve that is with the use of Web cookies. Governmental websites use cookies [16, 47], but it is not well studied if third-party cookies are also used when citizens visit such websites. Web cookies, also known as HTTP cookies, were introduced more than 30 years ago as a mechanism for websites to keep the state of a user's activity, e.g., recent visits and data entries, or for authentication. A Web cookie is a small piece of data stored on the user's computer by the Web browser when the user visits a website. The Web browser is in charge of handling cookies and storing them after begin created. Once a cookie is set, it is sent to the corresponding host, under the defined scope, along with each subsequent request until it is deleted or expired. Web cookies have been exploited to collect information about users' online activities and interests [12, 38].

In an attempt to put a stop to these profiling and tracking practices, new regulations mandate that the user has to be informed and give consent before cookies are stored on user's machine, e.g., the European Union General Data Protection Regulation (GDPR) [13] was put into effect on May 25, 2018. The GDPR levy fines against those who violate users' privacy and security standards, with penalties reaching twenty millions of euros or up to 4% of the annual worldwide turnover of the preceding financial year in case of an enterprise, whichever is greater. The regulation applies both to the private and the public sector and protects the rights of European citizens even when they visit websites outside European Union [3, 13, 21, 22, 25, 37]. Similar privacy regulations are now in effect also in other regions [24, 39], e.g., in California (California Consumer Privacy Act (CCPA) [44]), Canada [31], Israel [45], Japan [36], Australia [30], and Brazil [2].

WebSci '22, June 26–29, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9191-7/22/06.
https://doi.org/10.1145/3501247.3531545

Previous studies have demonstrated the widespread use of cookies for performing user tracking on the Web at an unprecedented scale [4, 25, 48]. Nevertheless, it is not well studied whether governmental websites enable similar cookie-based tracking, even unintentionally. Of course, one would expect that they do not, since these same governments are in charge of pushing anti-tracking initiatives via the above-mentioned laws. Things, however, are more complex. Oftentimes, third-party cookies sneak in inadvertently via the inclusion of links to social media and video portals or via "free" software modules and frameworks used to develop the website or service. Such software modules introduce their own cookies since they pursue business models based on tracking [35, 37, 42]. As a result, any website that relies on these external applications might unintentionally end up enabling tracking.

In this work, we inspect which cookies are set by governmental websites, with or without user consent, for how long, and by which tracking services. The *ramifications* of misusing Web cookies at governmental websites can be quite serious. First, it breaks the trust between citizens and authorities. Second, it allows for large-scale surveillance, monitoring, and tracking. If this takes place from third-parties it is worrisome as it shows bad website design that relies on external entities that can monitor interactions of the public with the government. Our objective in this paper is to shed light on those matters. The impact is more severe in case the website is the only point for interaction between the citizen and the government, or it is a lifeline resource for information, e.g., in the case of COVID-19 related official websites.

Our contributions can be summarized as follows:

- We perform a recent large-scale measurement study with more than 5.5k governmental websites and more than 118k URLs administrated by governments of countries worldwide, characterizing the ownership and time expiration of cookies added to website visitors.
- Contrary to our expectations and hope, we discover widespread tracking taking place between 9% up to 90% of the governmental websites of the twenty world's largest economies (G20) countries via tracking cookies that are added without user consent.
- More than 50% of cookies created on G20 government websites belong to third-parties and at least 10% (up to 90%) originate from known trackers. Most of these cookies have a life span of more than a day whereas many have an expiration time of a year or more.
- Our analysis also demonstrates a similar situation for official International organizations' websites, as well as for popular websites for COVID-19 related information.
- We make publicly available the datasets and the software we developed to enable future research on this topic. The data and the software are accessible at [20]: `https://govcookies.github.io/`

2 DATASETS

In this section we describe the websites selection, using publicly available sources, for (i) official governmental websites of G20 countries around the world, (ii) websites of International organizations, and (iii) popular official websites for COVID-19 related information.

Country	#Websites (Domains)	#Full URLs	#Unreachable URLs	Special top-level domains
Argentina	41	856	7	.gob.ar, .gov.ar
Australia	704	14,996	335	.gov.au
Brazil	494	10,558	1,093	.gov.br
Canada	116	2,474	108	.gc.ca
China	66	1,452	386	.gov.cn
France	243	5,198	1,333	.gouv.fr
Germany	226	4,884	136	-
India	1,429	30,000	5,124	.gov.in, .nic.in
Indonesia	48	1,032	41	.go.id
Italy	166	3,620	283	.gov.it
Japan	79	1,670	80	.go.jp
Mexico	118	2,527	299	.gob.mx
Russia	193	4,188	517	.gov.ru
Saudi Arabia	34	728	68	.gov.sa
South Africa	42	922	209	.gov.za
South Korea	42	896	55	.go.kr
Turkey	118	2,586	233	.gov.tr
UK	241	5,070	83	.gov.uk
USA	1,239	25,192	900	.gov

Table 1: Statistics for G20 Websites in our study.

2.1 G20 Governmental Websites

G20 is a group of governments and central bank governors from 19 countries and the European Union (EU). The 19 countries are listed in Table 1. For this study, we first present results for the 19 countries and later study the EU as an International organization. The G20 economies account for around 90% of the gross world product and two-thirds of the World population. The countries are located on different continents. Their cultural background and political standards vary; some are democracies, while monarchies or oligarchies govern others.

To compile the list of official websites for the G20 governments, we visit the official webpage of the government of each country and we collect all the links (URLs) to ministries and agencies that were listed there. For a list of sources that we used to collect these links, we refer the reader to [20]. In 18 out of these 19 countries, the websites use a special top-level domain, which helps us to validate that these websites are indeed the official ones. Nevertheless, many of the governmental websites use a different second-level domain than these special top-level domains. For a list of the special top-level domains for these countries, we refer to the last column in Table 1. An official website is a website associated to a domain that is registered and used by a national government. For example, `whitehouse.gov` is owned by the United States government and is used to release information about the current operations of the US President during his presidency. In Table 1 (second column) we report the overall number of official websites we identified. Note that these are typically the "landing" pages of the domain. As we will explain later (Section 3) for our study we visit multiple URLs in the same domain (Table 1, third column). In total, we consider 5,563 landing domains and 118,849 associated URLs.

A contemporary and independent study [40] considers 150,244 websites from 206 countries. Such study uses methodology to collect governmental websites that significantly differs from ours. Indeed, we focus on fewer countries and carefully investigate the presence of cookies at URLs provided by the official government websites.

Thus, a head-to-head comparison with the results presented in the mentioned study is not directly applicable.

2.2 International Organizations Websites

To compile the list of domains and URLs related to International organizations, we collect all the links of the agencies that are included in the official EU and UN websites. We also include in our list the major recognized international organization. For the list of sources we refer the reader to [20]. In total, our list includes 242 websites and 2,649 associated URLs.

2.3 COVID-19 Information Websites

We also compile a list of COVID-19 related URLs provided by the US Centers for Disease Control and Prevention (CDC) [18] and the European Center for Disease Prevention and Control (ECDC) [19] that provides links to official national sites in the EU. For each one of the G20 countries, we use SimilarWeb [43] to identify the most popular websites with the lemma COVID-19. In total, our list includes 131 official websites and 1,355 associated URLs related to COVID-19.

3 METHODOLOGY

In this section we review the types of Web cookies and we describe our methodology for crawling governmental and other official websites, including websites of International organizations and official websites related to COVID-19 pandemic.

3.1 Types of Cookies

First-party cookies are issued by the visited website, while third-party ones are typically created by external parties embedded in a webpage.The distinction is not always unambiguous, and in presence of *cookie ghostwriting* an entity creates cookies on behalf of another party [41]. Our goal is to provide a lower bound on third-party trackers, and to this end we focus only on those cookies which are not directly set by the visited domain. We perform an additional distinction on cookies using their *expiration time*. Session cookies are bound to the browser and once the browser process is terminated they get deleted. On the other hand, the lifetime of *persistent cookies* is set at their creation and they might last from few seconds up to several years.

3.2 Inferring Cookies

Usually cookies are created either via a *Set-Cookie* header in the server HTTP response, or they are set on the client side using JavaScript. We visit official websites with a modified version of *Pythia* [26], an open-source framework that instruments a fully-fledged browser to access URLs. Pythia was developed for analyzing hosting environments, and for this reason it does not provide information on cookies. We use the Chrome DevTools Protocol [5] to expand Pythia and collect *all the cookies* that are created when visiting a URL. Our approach is fully transparent to the browser and it keeps track also of cookies that are created by intermediate URLs which redirect the user to the landing page. After extracting all the cookies, we use the domains of the visited URLs to partition them into first and third-parties. Similarly, we inspect the expiration time of individual cookies and we label them either as session or persistent. We focus on the lifetime information, since previous studies [15, 25] considered as persistent any cookie with a lifetime of more than one day. Later, in our analysis, we provide statistics about the cookies lifetime and comment on the presence of persistent ones.

3.3 Identifying Tracking Cookies

We leverage filter lists to identify third-party cookies originating from known tracking services. Filter lists are an efficient solution to protect user's privacy by blocking ads and trackers. Two widely-used filter lists are Disconnect, core of Firefox tracking protection, and the Privacy Badger, maintained by the Electronic Frontier Foundation [8, 11, 28]. Other examples of list include EasyList, EasyPrivacy and Adguard [1, 9, 10]. These lists are manually curated and they offer protection only against trackers targeting the most popular services. We have to acknowledge, that being community-maintained projects, filter lists are often more suitable for western countries since they target popular services in these regions. Thus, in non-western countries the number of inferred third-party cookies by leveraging these lists is only a *lower bound*, as many more local third-parties not include in these list may also operate. We use the blocklists project [23] and we integrate it with the trackers extracted from Disconnect and Privacy Badger. Each filter list contains an index of domains and a set of rules that are used to detect services that harvest users' information. The rules are defined at the URL level, since it might happen that only a specific resource of a domain is responsible for tracking. From each list we extract the trackers domain names, making sure to select *only those domains that are fully blocked* by a particular filter list. Next, we enforce consensus among different lists and we flag as tracker any domain appearing in *two or more* filter lists. Those steps guarantee that (i) our curated list of trackers does not contain false positives, and that (ii) we select trackers which are both well-known and they appear across a wide range of services. In the final step, we label as *tracking cookie* any cookie that is set by a domain that is included in our list of popular trackers.

3.4 Crawling Websites

We bootstrap our analysis using the list of URLs that link to homepages of ministries and agencies, as well as the International organizations and COVID-19 webpages as described in Section 2. For all the URLs, we attempt to fetch the webpage using both HTTP and HTTPS as protocols, and we exclude and content which is retrieved with a status code associated to an error response (i.e., status code in the range 400-599). After extracting all the clickable hyperlinks from homepages, we discard all websites whose homepages contain less than two links. Next, we filter out links to different domains or to multimedia content (e.g., with extension ".jpg"), and we iteratively connect to each URL as we did with the homepages. During this step we inspect the MIME type of the retrieved content, and we discard any resource which does not contain HTML. We repeat this process until we have collected information from 10 unique HTML resources on a particular website. We sample only a small number of internal webpages both to avoid stressing the server and to guarantee consistency among websites (e.g., some websites might host only a dozen webpages, while others might have thousands).

Table 1 (third column) provides an overview of the overall number of governmental and COVID-19 websites, together with the corresponding number of URLs that we accessed. To our surprise, we identify 12,623 (around 11% of the overall) URLs related to the nineteen G20 countries, 9 URLs related to International organizations, and 3 URLs related to COVID-19 that were unreachable, see Table 1 (last column). Manual investigation shows that indeed the URL links were broken or the URLs were offline or have been removed during our visits (see Section 4 for details about our experiments).

Once we compiled a set of URLs associated to each website, we use our framework to visit each URL and collect the cookies. We access each URLs running the Google Chrome with Pythia on a Linux/Debian desktop, with the browser configured to accept all third-party cookies. Our instrumented browser never interacted with the loaded webpages (e.g., scrolling or moving the mouse pointer) nor did it perform any action that could be interpreted as consent granting (e.g., clicking a dialog box/prompt/pop-up). Here we note that the ePrivacy Directive [14] requires explicit user consent for cookie tracking.

We collect the data from two countries in the European Union, thus, we would expect that no cookies should be installed because there is no action taken by the instrumented client. Our results confirm previous studies that show that indeed cookies are installed without the user giving consent or even when the user explicitly choose the option to reject all cookies [7, 42, 46, 49]. Moreover, previous studies have found that having cookies set in the browser increases by 27% the amount of third-party cookies that are observed [48]. Our goal is to provide a *lower bound* on cookies that are served, and for this reason we access each URL as new user that is accessing the website for the first time. To mimic a new visitor, before loading each URL, we configure our framework to delete previously stored cookies, the browser history, the cache and the local storage and disable all extensions.

3.5 Ethical Considerations

We do not use or collect any personal data or real users to perform our experiments. Data collection was done using an instrumented client from universities in two EU countries. We also scheduled the experiments such that the load on the servers of governmental, international organizations, and COVID-19 related websites to be minimal, i.e., at most one visit per domain per minute. For the duration of our experiments, we use the same IPv4 addresses for all our experiments. During and after our experiments, we did not receive any complaints by network centers of the hosting universities or the administrators of the governmental, International organizations, and COVID-19 related websites we visited.

4 ANALYSIS

Our instrumented client visited the G20 governmental URLs in October 2020, and the International organizations and COVID-19 URLs in March 2021. We first analyze in detail the cookies that are installed during our visits to the G20 governmental websites and then repeat the same for the International organizations and the COVID-19 websites.

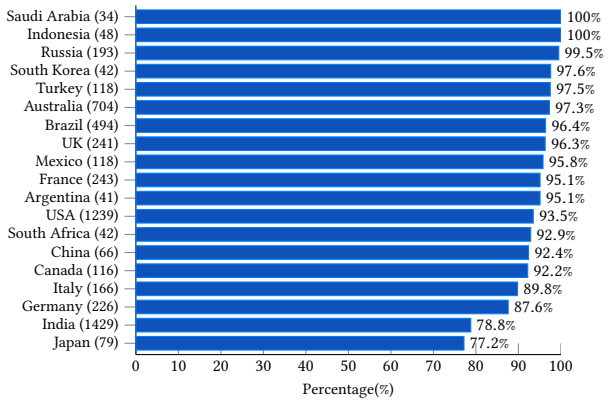

Figure 1: Percentage of government websites (number in parenthesis) that contain ≥ 1 cookie per G20 country.

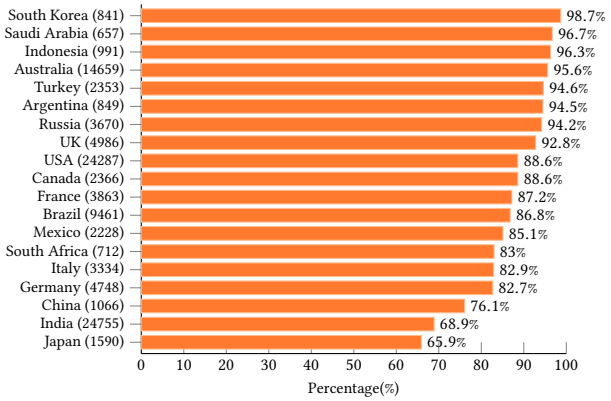

Figure 2: Percentage of government URLs (number in parenthesis) that contain ≥ 1 cookie per G20 country.

4.1 G20 Websites

In Figure 1, we present the percentage of websites per country that install at least one cookie. In parenthesis we include the number of websites that we visited per G20 country. The majority of the official websites, ranging from 77% to 100%, of the G20 countries indeed add cookies, without any user consent. To confirm that our results are not biased due to the contributions of a single URL, we inspect the aggregated results across all of the URLs. In Figure 2 we group the URLs per G20 country, and we show the overall percentage of URLs with *at least one cookie*. For each country, in parenthesis we report the total number of URLs that were visited. A comparison among Figures 1 and 2 suggests that percentages decrease only slightly when URLs that belong to the same website are not grouped together. We notice also some small fluctuations in the ranking of countries. The general observation, however, remains as the large majority of URLs set cookies, ranging from 68% to more than 95%.

These probabilities are significantly higher than previous studies conducted in 2019 that report that only 15%-50% of websites in the category "Law and Government" set cookies [42, 46]. However, these previous studies do not explicitly consider governmental websites only, as the category "Law and Government" refer to sites tagged as such by advertisement companies, such as Google Ads.

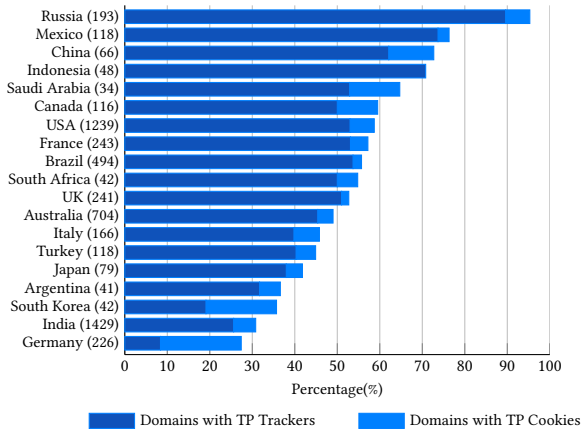

Figure 3: Percentage of government websites with third-party (TP) and third-party tracker (TPT) cookies per G20 country.

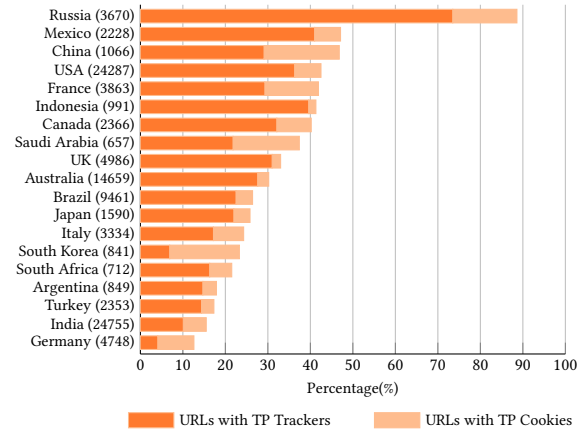

Figure 4: Percentage of URLs that contain third-party (TP) and third-party tracker (TPT) cookies.

Since our list of domains is more recent and compiled in a different way (see Section 3), a head-to-head comparison with these may be misleading.

The percentage of domains and URLs related to the G20 countries with at least one cookie is very high. To put these percentages in perspective, we compare with other studies that studied millions of URLs. Each one of the methods follows a slightly different methodology, and the user population varies from mobile users [35] to users that browse the Web with a specific browser [50]. The number of visited websites also differs. Nevertheless, all the studies agree that around 85%-95% of the visited URLs add at least one first- or third-party cookie. Other studies also showed that cookies are added even when users (from different parts of the world) do not give their consent [7, 21, 22, 35, 38, 42, 46, 49]. Our analysis shows that G20 websites are not an exception. Thus, in general, no special care has been taken when designing governmental websites. It is also striking that it is more likely for a visitor to receive cookies when visiting around a third of G20 countries than the highest reported number by studies that consider general websites.

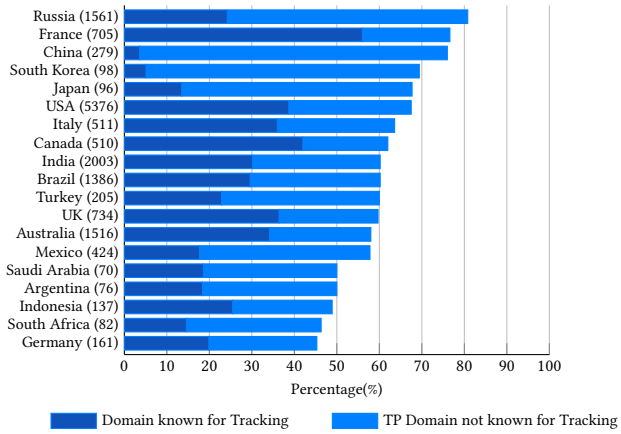

Figure 5: Percentage of TP and third-party trackers (TPT) cookies with expire times ≥ a day for G20 countries.

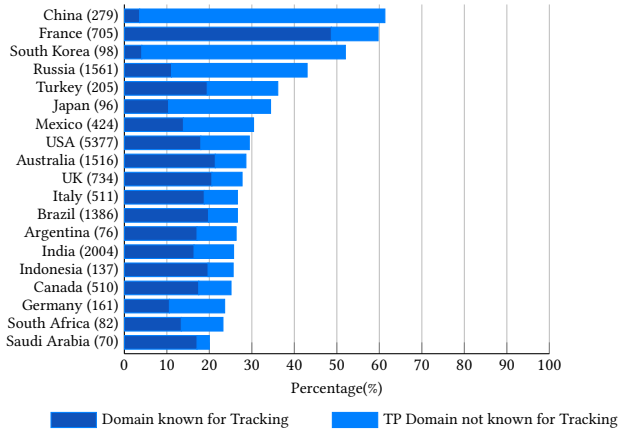

Figure 6: Percentage of TP and third-party trackers (TPT) cookies with expire times ≥ a year for G20 countries.

In terms of the number of cookies set by first and third-parties, the number varies a lot across countries. For the majority of the countries, the average number of cookies set at each visit is less than 8. This number is lower than the corresponding average number of cookies, i.e., 12 reported by a recent study for general websites [35]. However, there are some exceptions. Governmental websites of Russia and China typically set 12 or more cookies with about half of them associated with third-parties. We elaborate more on the role of third-party cookies and trackers in the following sections.

4.1.1 Third-party Cookies. We then turn our attention to the type of cookies that are added when we visit an official webpage or URL. Someone would argue that first-party cookies may be used to optimize the user experience, however, no privacy expert would advocate in favor of adding third-parties (TP) cookies on an official governmental website. As shown in Figure 3, a large fraction of official websites add third-party cookies. This observation applies across the board. The percentage of websites that add at least one third-party cookie in Russia is similar to the percentage of websites that add any cookie. Indeed, half of the domains in ten other G20 countries add third-party cookies. Even when we study Germany,

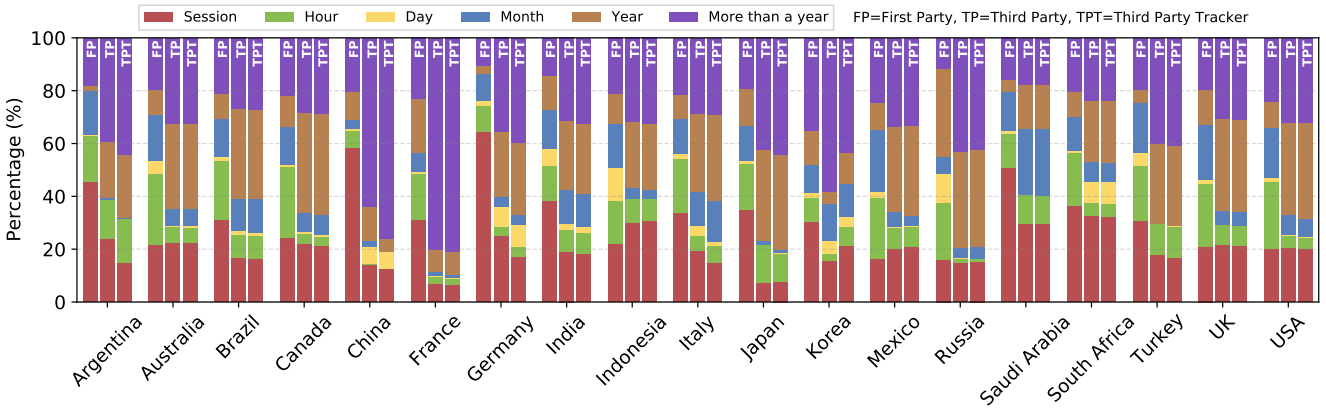

Figure 7: Expiration times for first-party (FP), third-party (TP), and third-party trackers' (TPT) cookies at G20 countries.

a country known for its strict user privacy regulation, we notice that more than 25% of the official websites add third-party cookies.

In Figure 4, we repeat the same analysis for the URLs of official webpages from the G20 countries, but this time focusing on third-party cookies. We can observe that percentage of URLs that set third-party cookies in this case reduces. However, this does not seemt o affect the ranking. Around 90% of the URLs in official Russian websites are adding third-party cookies. Moreover, more than 30% of URLs of official websites of other eight G20 countries add at least one third-party cookie.

4.1.2 Tracking. Next we focus on third-party tracker (TPT) cookies, i.e., cookies set by domains that are known to be tracking users for data collection purposes (see also Section 3). In Figure 3, we annotate with dark blue color the fraction of websites that add at least one cookie associated with a tracking domain. These percentages are only slightly lower than the corresponding ones for third-party cookies in general for the same G20 country. Germany is the only country where this percentage decreases significantly, and only 9% of the official websites include a cookie from a tracking domain. Similar observations apply at the URL level, as shown in Figure 4. Indeed, the percentage of URLs that add cookies of tracker domains (annotated with dark orange color) is lower but comparable to those that add third-party cookies in general, with the noticeable exceptions of Germany and South Korea. A related study is a report by CookieBot in 2019 that only considered one official landing page for each European Union country [6]. It found that 89% of these domains contain third-party ad tracking. Our analysis, on a much larger set of landing pages, shows a lower presence of trackers with significant variation (5% to 30%) in government domains across EU countries. Apart from the existence, or lack of, tracking, another important aspect has to do with its severity. The latter is more pronounced when tracking cookies have a long lifetime [7]. In Figure 5 we report the percentage of the cookies of third-parties (light blue) that expire in more than one day, as previous studies characterize such cookies as *persistent* cookies [15, 25]. The values in parenthesis report the total number of cookies added when we visited the URLs associated with official websites in each G20 country. The percentages are very high across all countries, with more than 50% for the majority of the cookies added in websites in G20 countries after more than one day. To put these percentages in perspective, a previous study that run experiments on 35k domains of general interest reports that for around 85% of the third-party cookies has a lifetime of one day or more [46]. Thus, we conclude that the percentage of third-parties with lifetime more than a day is high, but typically lower than the average in general webpages. There are three exceptions, Russia, France, and China where around 75% of the cookies set by third-parties last for at least one day.

In Figure 6 we also report the percentage of cookies set by third-parties with lifetime of one year or more. A significant percentage, between 20% to 60% depending on the country, of the cookies set by third-parties last for a year or more. For the majority of the G20 countries, more than 25% of the added cookies related to third-parties last for a year or more. Typically, the percentage is lower than the percentage (50%) of cookies that last for a year or more and are set by third-parties when accessing general websites [46]. However, for three countries, namely China, France, and South Korea, the percentages of third-party cookies with lifetime a year or more is higher than the percentage in general websites.

When we consider the cookies that are associated with known trackers (dark blue) as we annotate them based on our methodology presented in Section 3, the percentages are lower, but still significant. Indeed, between 5% to 55% depending on the country, of the cookies set by trackers last for a day or more, see Figure 5. Countries like France (55%), Canada (42%), US and UK (both around 35%), and Italy and Australia (both around 30%) have persistent cookies set by trackers and last for more than one day. Even more striking is the observation that there are third-party and tracking cookies with expiration time longer than a year in governmental websites of G20 countries. In Figure 6, we report these percentages. A significant percentage, between 5% to 50% depending on the country, of the cookies set by trackers and last for a year or more. For the majority of the G20 countries, more than 25% of the added cookies related to third-parties last for a year or more. Nevertheless, even in two countries with strong protection regulation such as Germany and France there are many tracking cookies with expiration times beyond one year.

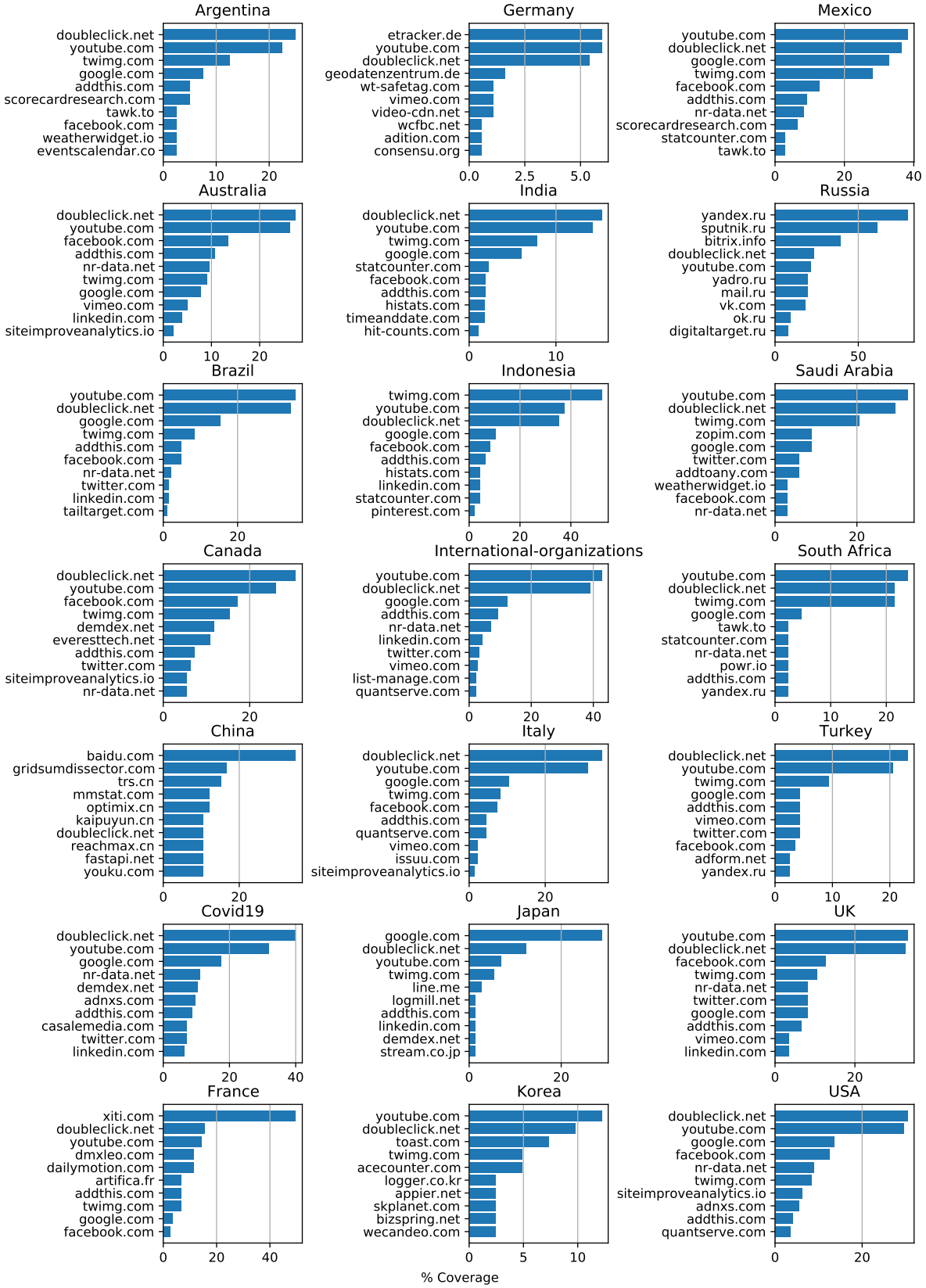

Figure 8: Popular third-party trackers in G20 countries, International organizations, and COVID-19 websites. X-axis: % coverage.

Country	Domain	# of Trackers	Trackers
Argentina	munirivadavia.gob.ar	5	doubleclick.net, youtube.com, microsoft.com, office.com, twimg.com
Australia	sea.museum	13	doubleclick.net, adnxs.com, casalemedia.com, adroll.com, outbrain.com, bidswitch.net, yahoo.com, advertising.com, pubmatic.com, 3lift.com, taboola.com, facebook.com, openx.net
Brazil	investexportbrasil.gov.br	25	vertamedia.com, relap.io, rktch.com, mail.ru, datamind.ru, bumlam.com, betweendigital.com, crwdcntrl.net, upravel.com, digitaltarget.ru, doubleclick.net, 1dmp.io, mts.ru, advarkads.com, aidata.io, rutarget.ru, uuidksinc.net, acint.net, yandex.ru, republer.com, adsniper.ru, beroll.ru, sape.ru, adriver.ru, adhigh.net
Canada	nac-cna.ca	25	demdex.net, admanmedia.com, adnxs.com, exelator.com, casalemedia.com, 3lift.com, lijit.com, advertising.com, pro-market.net, crwdcntrl.net, bing.com, yahoo.com, doubleclick.net, openx.net, acuityplatform.com, youtube.com, smartadserver.com, tapad.com, smaato.net, addthis.com, sonobi.com, adsrvr.org, facebook.com, pubmatic.com, bidswitch.net
China	stats.gov.cn	9	doubleclick.net, reachmax.cn, gridsumdissector.com, fastapi.net, chinavivaki.com, admaster.com.cn, youku.com, optimix.cn, trs.cn
France	service-civique.gouv.fr	24	demdex.net, quantserve.com, everesttech.net, google.com, avct.cloud, adnxs.com, casalemedia.com, krxd.net, rfihub.com, advertising.com, adotmob.com, serving-sys.com, rlcdn.com, xiti.com, bing.com, doubleclick.net, yahoo.com, bluekai.com, eyeota.net, media.net, rezync.com, facebook.com, spotxchange.com, bidswitch.net
Germany	bund.de	5	doubleclick.net, appdomain.cloud, youtube.com, geodatenzentrum.de, vimeo.com
India	kerala.gov.in	7	doubleclick.net, google.com, twimg.com, tawk.to, youtube.com, addthis.com, facebook.com
Indonesia	big.go.id	5	pinterest.com, youtube.com, google.com, linkedin.com, twimg.com
Italy	www.difesa.it	4	doubleclick.net, google.com, youtube.com, twimg.com
Japan	jpf.go.jp	4	logmill.net, doubleclick.net, twimg.com, google.com
Korea	moef.go.kr	6	logger.co.kr, appier.net, toast.com, twimg.com, skplanet.com, bizspring.net
Mexico	acapulco.gob.mx	6	doubleclick.net, weatherwidget.io, youtube.com, addthis.com, google.com, twimg.com
Russia	gov.ru	31	kitbit.net, semantiqo.com, yadro.ru, google.com, ok.ru, uptolike.com, rktch.com, mail.ru, bumlam.com, twimg.com, pluso.ru, sputnik.ru, bitrix.info, upravel.com, digitaltarget.ru, doubleclick.net, trum-trum.club, pinterest.com, youtube.com, aidata.io, rutarget.ru, yandex.ru, adsniper.ru, konverbot.com, nr-data.net, weborama.fr, cdnvideo.ru, caltat.com, facebook.com, yandex.com, vk.com
Saudi Arabia	alqassim.gov.sa	4	doubleclick.net, weatherwidget.io, twimg.com, youtube.com
South Africa	dtps.gov.za	4	doubleclick.net, twimg.com, youtube.com, statcounter.com
Turkey	botas.gov.tr	9	squareup.com, reddit.com, pinterest.com, expedia.de, foursquare.com, google.com, twitter.com, tumblr.com, dropbox.com
UK	startuploans.co.uk	7	doubleclick.net, pardot.com, force.com, salesforceliveagent.com, youtube.com, bing.com, facebook.com
USA	hhs.gov	13	doubleclick.net, demdex.net, turn.com, intentiq.com, adentifi.com, sc-static.net, youtube.com, mxptint.net, yahoo.com, quantserve.com, twitter.com, snapchat.com, facebook.com

Table 2: G20 government related domains per country with the highest number of third-party trackers (TPT) in our study and associated trackers.

Overall, a common characteristic across the board is that the cookies set by trackers expire later than the cookies set by non-tracking third-parties and first-parties. Moreover, they rarely expire in less than one hour or after the end of the session. To investigate this further, in Figure 7 we plot the expiration times for first-, third-party, and third-party trackers cookies at G20 countries' government websites. The majority of cookies set by third-party trackers lasts for a month or more, which is a significantly higher percentage when compared with the percentage of first- and third-party cookies, respectively. Note that long-lasting cookies allow trackers to gather much more data about website visitors.

4.1.3 Profiling Trackers. Finally, we investigate why there are so many trackers present in governmental websites. To shed light, we study the ten most popular trackers in governmental websites for each country. We define a metric that we call *coverage*, as the percentage of governmental websites (domains) in a country where a given tracker is present. We plot our results in Figure 8. A first observation is that almost all G20 countries trackers operated by Google (`doubleclick.com`, `youtube.com`, `google.com`) related to analytics are at the very top, with a coverage between 20% and 50%. A noticeable exception is China, where only one of the Google trackers is present in the top 10 list, namely, `doubleclick.com`. In this case the coverage of this single tracker is around 10%. The tracker of social network Facebook (`facebook.com`) is present in 14 out of 19 G20 countries, but typically the coverage is lower than Google, between 5% to 20% depending on the country. The tracker of social network LinkedIn (`linkedin.com`) follows in popularity with presence in 7 of the G20 countries. It is worth mentioning that we observe regional trackers that are very popular in some countries. For example, the XiTi tracker (`xiti.com`) in France with coverage around 50%, the Baidu tracker (`baidu.com`) in China with coverage around 40%, the tracker of social network Twitter (`twimg.com`) with coverage close to 50% in Indonesia, and tracker of analytics `yandex.ru` and the social network V Kontakte `vk.com` with coverage more than 60% and 20%, respectively, in Russia.

Our manual investigation shows that many of these trackers are added because many of the governmental sites include links to social networks such as Facebook and LinkedIn and link videos hosted on Youtube or Vimeo. Another reason for the high presence of trackers is that many governmental web pages use analytics tools to monitor the number of their visitors. Popular social networks and video hosters offer such analytics tools, search engines like Google and Yandex, and smaller web traffic analysis companies, e.g., XiTi (France). In fewer cases, the designer of the webpages utilize Web libraries, e.g., of Google that can act as trackers. Unfortunately, our analysis shows that there is no apparent provision to remove third-party trackers altogether from official governmental websites, as we would hope and expect from administrators of such websites.

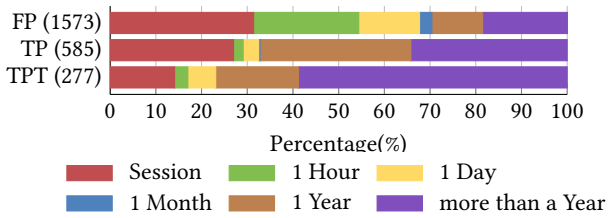

Figure 9: Expiration times for first-, third-party, and trackers (TPT) cookies at International organizations' domains.

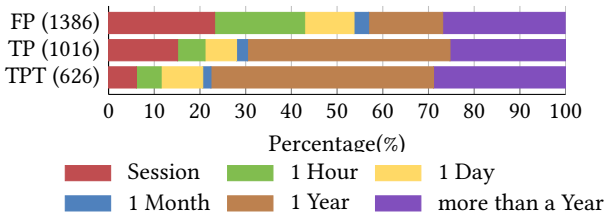

Figure 10: Expiration times for first-, third-party (TP) , and trackers (TPT) cookies at COVID-19 websites.

Hostname	#Trackers
www.wfp.org	36
icsid.worldbank.org	13
www.itu.int	13
www.worldfishcenter.org	13
irena.org	12
www.irena.org	12
www.worldbank.org	12
www.glfc.org	10
www.miga.org	10
www.adb.org	9

Table 3: Top 10 International organizations by trackers.

Hostname	#Trackers
coronavirus.jhu.edu	7
www.landlaeknir.is	6
covid19.min-saude.pt	4
www.spkc.gov.lv	4
deputyprimeminister.gov.mt	3
eody.gov.gr	3
koronavirus.gov.hu	3
www.cdc.gov	3
www.folkhalsomyndigheten.se	3
www.gouvernement.fr	3

Table 4: Top 10 official COVID-19 websites by trackers.

We also identify popular governmental websites with a very high number of trackers. In Table 2 we list, for each country, the domains with the highest number of trackers and the associated trackers that set cookies without any user consent. In some of these official governmental domains, tens of trackers are indeed present. Manual investigation shows that, again, the majority of the trackers are related to analytics and social network companies (local or international ones). Some of the domains with a high number of trackers are public broadcasters, e.g., `dw.com`, `sbs.com.au`.

4.2 International Organizations Websites

We also study official websites of International organization (see Section 2 for details). Our analysis shows that around 95% of the International organizations websites set cookies and around 60% of these websites use at least one third-party cookie. These percentages are close to these reported for general websites [46]. Thus, it seems there is no special care not to neutralize third-party cookies either in these websites. Around 52% of the International organizations websites set at least one cookie associated with a tracker. In Figure 9 we show the expiration time for first-party, third-party, and trackers cookies. The values in the parenthesis are the total number of cookies for each category. We note that the fraction of cookies that expire in more than one day, i.e., considered persistent, is 85% and 68% for tracker and third-party cookies respectively, which is way higher than this of first-party cookies that are around 45%. In the case of trackers, more than 75% of the cookies they set expires in a year or more. Thus, the presence of trackers in these websites raises serious concerns about the privacy of its visitors as these web cookies are stored for an extended period.

In terms of numbers of cookies set by the webpage, the average number is close to 12, as reported for general websites [35]. However, when we focus only on trackers we noticed that there are popular official websites with a very high number of trackers. In Table 3 we list the top 10 International organizations websites with the number of trackers that set cookies in the website without any visitor consent. It is striking that up to tens trackers are present, with the recently Nobel-awarded World Food Program to host 36 trackers and World Bank as well as the International Telecommunication Union (a United Nations specialized agency) websites to have 13 trackers present. As we show in Figure 8, popular trackers operated by Google have coverage of about 60% and other trackers operated by Twitter, LinkedIn, and Vimeo are also present.

4.3 COVID-19 Websites

As a final case study, we turn our attention to websites that provide information about COVID-19. Our analysis shows that more than 99% of these websites add at least one cookie without consent. This percentage is even higher than the one reported for general websites [35]. We compare our results with these from a study in May 2020 that found that more than 99% of websites that are returned when submitting Google queries related to COVID-19 add cookies of third-parties [27]. We observe a smaller presence of third-party cookies, around 62%. We argue that this is difference is due to the different set of websites we consider and the approach we sue to identify those websites. For instance, many of the search results are commercial websites that sell ad space to advertisers and not official COVID-19 information websites. Our analysis also shows that more than half of these websites add at least 3 third-party cookies. Moreover, almost all the websites with third-party cookies,

around 60%, also set tracking cookies. In Figure 10 we show the expiration time for first-party, third-party, and trackers cookies. The values in the parenthesis are the number of cookies for each category. We notice that the fraction of cookies that expire in a day or more is 95% and 85% for trackers and third-party cookies, respectively. These percentages are way higher than the corresponding of first-party cookies that are around 60%. Moreover, around 78% of the trackers cookies last for a year or more.

Finally, we report on COVID-19 websites with the most trackers. Our analysis shows that news websites have the highest number of trackers, some of them with more than 30 trackers. This is to be expected as the business model of the news sites is to attract user visits and advertisers. However, when we focus on the official COVID-19 related websites that are operated by international or national health organizations and governments, we notice that the number of trackers that set cookies, without any consent, is also high (see Table 4). For example, the very popular website with global maps about the COVID-19 cases, maintained by Johns Hopkins University, add cookies from 7 trackers. All the other Top 10 website are official national information websites in European countries that have three trackers or more. The American Centers for Disease Control and Prevention (CDC) is also in the Top 10, with cookies associated with three trackers. Trackers operated by Google are present in more than half of COVID-19 related websites, as shown in Figure 8. Other popular trackers operated by social media, e.g., Twitter and LinkedIn, are also present in around 10% of these websites. We conclude that the presence of trackers in websites related to COVID-19 that are very popular during the pandemic is high and, typically, there is no special care to remove them.

5 DISCUSSION

Responsible Governmental Website Development. Our study shows the critical role that official websites play in informing citizens, but also the potential privacy harm due to third-party tracking. The designers and contractors of official websites, e.g., government websites and websites related to health, need to take extra care to (i) avoid including plugins for social media, commercial video portals, and publishers (advertising media), and (ii) avoid including links that download content from such websites. If there is a need to add a link to social media, a good practice is to add a logo, inform the visitors that they are leaving the official website when they click on the logo, and redirect to the social media. It is also important to use software and libraries that have been certified and do not leak private information about website visitors. Finally, the software companies and contractors need to self-regulate the industry when it comes to protecting visitors of sensitive sites related to government and health.

Governmental Cloud. Our study shows that trackers operated by video portals, social media, and analytics companies are among the most popular. A possible solution to this problem is a government-owned and operated cloud that hosts and delivers videos and content to citizens. Although this solution is more expensive and requires additional investment in human capital and expertise, it is more sustainable in long run.

Governmental Websites Audit. Independent authorities in each country should perform regular and detailed audit campaigns to assess the state of third-party tracking in governmental and health-related websites. Frequent audits should report on third-party tracking and swiftly remove trackers from such websites. Civil Societies and researchers can also perform independent audits and disclosure to authorities about user tracking in governmental and websites of public interest. The tools that we release in this paper can be used to help in the auditing process.

Education about Web Tracking. It is important to increase awareness of the public about the potential harm of user tracking. Teenagers should learn more about Web technologies and the shortcomings of tracking at School. Professionals should also attend seminars to become familiar with the downsides of user tracking. Research communities and civil societies can also organize events and hackathons to raise awareness about tracking and contribute to adopting the best current practices to reduce tracking on public and health-related websites.

6 CONCLUSION

In this paper, we present a recent large-scale measurement study of cookie presence at governmental sites and other popular non-commercial sites that have high visibility with the public. Our focus is on third-party cookies and well known tracking services. Ironically, it seems that despite great efforts to promote regulations like GDPR, governmental sites themselves, are not yet clear of tracking practices targeted by such regulations. Our results indicate that official governmental, international organizations' websites and other sites that serve public health information related to COVID-19 are not held to higher standards regarding respecting user privacy. Our analysis shows that trackers are widely present at such websites, and cookies are added without user consent as developers or administrators of these websites, , probably unintentionally, include external content from social media and third-party services. Our work demonstrates how difficult it is to apply data protection laws in practice, and we hope that it can help in clearing governmental websites and similar webpages that serve public services from tracking services. With our study, we also aim to increase awareness of potential tracking when visiting official websites, and we argue for the need for new tools and systems for continuous measurement and transparent reporting to improve the privacy of public online services.

ACKNOWLEDGMENTS

This work was supported in part by the Atracción de Talento grant (Ref. 2020-T2/TIC-20184) funded by Madrid regional government, the TV-HGGs project (OPPORTUNITY/0916/ERC-CoG/0003) co-funded by the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation, the European Research Council (ERC) Starting Grant ResolutioNet; "Resolving the Tussle in the Internet: Mapping, Architecture, and Policy Making" (ERC-StG-679158), and the EU H2020 Research and Innovation programme under grant agreements No. 871370 (Pimcity).

REFERENCES

[1] AdGuard. 2021. AdGuard – World's most advanced adblocker!. https://adguard.com/.

[2] Brazil. 2018. Lei Geral de Proteção de Dados. https://iapp.org/media/pdf/resource_center/Brazilian_General_Data_Protection_Law.pdf.

[3] C. Iordanou and G. Smaragdakis and I. Poese and N. Laoutaris. 2018. Tracing Cross Border Web Tracking. In *IMC*.

[4] A. Cahn, S. Alfeld, P. Barford, and S. Muthukrishnan. 2016. An Empirical Study of Web Cookies. In *WWW*.

[5] Google Chrome. 2021. Chrome DevTools Protocol. https://chromedevtools.github.io/devtools-protocol/.

[6] CookieBot. 2019. Ad Tech Surveillance on the Public Sector Web. https://www.cookiebot.com/media/1121/cookiebot-report-2019-medium-size.pdf.

[7] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz. 2019. We Value Your Privacy ... Now Take Some Cookies - Measuring the GDPR's Impact on Web Privacy. In *NDSS*.

[8] Disconnect. 2021. Best privacy VPN app for iOS and Mac. Powerful protection with one tap. https://disconnect.me/.

[9] EasyList. 2021. EasyList - Overview. https://easylist.to/.

[10] EasyPrivacy. 2021. EasyPrivacy list. https://easylist.to/easylist/easyprivacy.txt.

[11] Electronic Frontier Foundation. 2021. Privacy Badger. https://privacybadger.org/.

[12] S. Englehardt, D. Reisman, C. Eubank, P. Zimmerman, J. Mayer, A. Narayanan, and E. W. Felten. 2015. Cookies That Give You Away:The Surveillance Implications of Web Tracking. In *WWW*.

[13] European Commission. 2018. Data protection in the EU, The General Data Protection Regulation (GDPR); Regulation (EU) 2016/679. https://ec.europa.eu/info/law/law-topic/data-protection/.

[14] European Commission. 2019. Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009. Official Journal of the European Union. (https://eurlex.europa.eu/legal-content/EN/TXT/?uri=celex:32009L0136).

[15] T. Favale, F. Soro, M. Trevisan, I. Drago, and M. Mellia. 2020. Campus Traffic and e-Learning during COVID-19 Pandemic. https://arxiv.org/abs/2004.13569.

[16] US federal government. 2021. Privacy and Security Policies. https://www.usa.gov/policies.

[17] A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, C. Dietzel, D. Wagner, M. Wichtlhuber, J. Tapiador, N. Vallina-Rodriguez, O. Hohlfeld, and G. Smaragdakis. 2021. A Year in Lockdown: How the Waves of COVID-19 Impact Internet Traffic. *Communications of the ACM* 64, 7 (2021).

[18] US Centers for Disease Control and Prevention. 2021. Coronavirus Disease 2019. https://www.cdc.gov/coronavirus/2019-ncov/index.html.

[19] European Centre for Disease Prevention and Control. 2021. External resources on COVID-19. https://www.ecdc.europa.eu/en/covid-19/external-resources.

[20] M. Götze, S. Matic, C. Iordanou, G. Smaragdakis, and N. Laoutaris. 2022. Artifacts of the paper: "Measuring Web Cookies in Governmental Websites", ACM WebSci'22. https://govcookies.github.io/.

[21] X. Hu and N. Sastry. 2019. Characterising Third Party Cookie Usage in the EU after GDPR. In *WebSci*.

[22] J. Sorensen and S. Kosta. 2019. Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. In *WWW*.

[23] JustDomains. 2021. GitHub - justdomains/blocklists: Domain-ONLY Filter Lists (for use with DNS / Domain blocking tools). https://github.com/justdomains/blocklists.

[24] L. Kalman. 2019. New European Data Privacy and Cyber Security Laws: One Year Later. *CACM* 62, 4 (2019).

[25] S. Matic, C. Iordanou, G. Smaragdakis, and N. Laoutaris. 2020. Identifying Sensitive URLs at Web-Scale. In *IMC*.

[26] S. Matic, G. Tyson, and G. Stringhini. 2019. PYTHIA: a Framework for the Automated Analysis of Web Hosting Environments. In *WWW*.

[27] M. S. McCoy, T. Libert, D. Buckler, D. T. Grande, and A. B. Friedman. 2020. Prevalence of Third-Party Tracking on COVID-19–Related Web Pages. *Jama* 324, 14 (2020).

[28] Mozilla. 2019. Today's Firefox Blocks Third-Party Tracking Cookies and Cryptomining by Default - The Mozilla Blog. https://blog.mozilla.org/blog/2019/09/03/todays-firefox-blocks-third-party-tracking-cookies-and-cryptomining-by-default/.

[29] United Nations Department of Economic and Social Afairs. 2021. Digital Government. https://publicadministration.un.org/en/ict4d.

[30] Office of the Australian Information Commissioner. 2018. Australian Privacy Principles guidelines; Australian Privacy Principle 5 – Notification of the collection of personal information. https://bit.ly/38BBrUP.

[31] Office of the Privacy Commissioner of Canada. 2018. Amended Act on The Personal Information Protection and Electronic Documents Act. https://bit.ly/3oPDSJ7.

[32] Organisation for Economic Co-operation and Development. 2008. Implementing E-governance in OECD Countries: Experiences and Challenges. https://www.oecd.org/mena/governance/36853121.pdf.

[33] Organisation for Economic Co-operation and Development. 2021. Productivity Growth in the Digital Age. https://www.oecd.org/going-digital/productivity-growth-in-the-digital-age.pdf.

[34] Organisation for Economic Co-operation and Development. 2021. Responding to Covid-19: The rules of good governance apply now more than ever! https://www.oecd.org/governance/public-governance-responses-to-covid19/.

[35] P. Papadopoulos, N. Kourtellis, and E. Markatos. 2019. Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. (2019).

[36] Personal Information Protection Commission, Japan. 2017. Amended Act on the Protection of Personal Information. https://www.ppc.go.jp/en/.

[37] A. Razaghpanah, R. Nithyanand, N. Vallina-Rodriguez, S. Sundaresan, M. Allman, C. Kreibich, and P. Gill. 2018. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *NDSS*.

[38] S. Englehardt and A. Narayanan. 2016. Online Tracking: A 1-million-site Measurement and Analysis. In *ACM CCS*.

[39] S. Greengard. 2018. Weighing the Impact of GDPR. *Comm. of the ACM* 61, 11 (2018).

[40] N. Samarasinghe, A. Adhikari, M. Mannan, and A. Youssef. 2022. Et tu, Brute? Privacy Analysis of Government Websites and Mobile Apps. In *WWW*.

[41] I. Sanchez-Rola, M. Dell'Amico, D. Balzarotti, P.-A. Vervier, and L. Bilge. 2021. Journey to the Center of the Cookie Ecosystem: Unraveling Actors' Roles and Relationships. In *IEEE Symposium on Security and Privacy*.

[42] I. Sanchez-Rola, M. Dell'Amico, P. Kotzias, D. Balzarotti, L. Bilge, P.-A. Vervier, and I. Santos. 2019. Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control. In *ACM AsiaCCS*.

[43] SimilarWeb. 2021. Coronavirus Data, Insights & Trends (Covid-19). https://www.similarweb.com/coronavirus/.

[44] State of California. 2018. California Consumer Privacy Act – Assembly Bill No. 375. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

[45] The Privacy Protection Authority of Israel. 2018. Protection of privacy regulations (data security) 5777-2017. https://bit.ly/2LMwcIA.

[46] M. Trevisan, S. Traverso, E. Bassi, and M. Mellia. 2019. 4 Years of EU Cookie Law: Results and Lessons Learned. *PETS* 2019, 2 (2019), 126–145.

[47] UK government. 2021. Cookies on GOV.UK. https://www.gov.uk/help/cookies.

[48] T. Urban, M. Degeling, T. Holz, and N. Pohlmann. 2020. Beyond the Front Page: Measuring Third Party Dynamics in the Field. In *WWW*.

[49] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *ACM CCS*.

[50] Z. Yu, S. Macbeth, K. Modi, and J. M. Pujol. 2016. Tracking the Trackers. In *WWW*.