Accepted Manuscript

Towards a Global Perspective on Web Tracking

Nayanamana Samarasinghe, Mohammad Mannan

PII: S0167-4048(18)31400-7
DOI: https://doi.org/10.1016/j.cose.2019.101569
Article Number: 101569
Reference: COSE 101569

To appear in: *Computers & Security*

Received date: 30 November 2018
Revised date: 11 July 2019
Accepted date: 11 July 2019

Please cite this article as: Nayanamana Samarasinghe, Mohammad Mannan, Towards a Global Perspective on Web Tracking, *Computers & Security* (2019), doi: https://doi.org/10.1016/j.cose.2019.101569

Towards a Global Perspective on Web Tracking

Nayanamana Samarasinghe*, Mohammad Mannan

Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada

Abstract

Several past measurement studies uncovered various aspects of web-based tracking and its serious impact on user privacy. Most studies used institutional resources, e.g., computers hosted at well-known universities, or cloud-computing infrastructures such as Amazon EC2, confining the study to a particular geolocation or a few locations. Would there be any difference if web tracking is measured from actual user-owned residential machines? Does a user's geolocation affect web tracking? Past studies do not adequately answer these important questions, although web users come from across the globe, and tracking primarily targets home users. As a step forward, we leverage the Luminati proxy service to run a measurement study using residential machines from 56 countries. We rely on the OpenWPM web privacy measurement framework to analyze third-party scripts and cookies in 2050 distinct URLs (Alexa Top-1000 home pages and Alexa Top-50 country-specific home pages for all 56 countries, and shared URLs via Twitter from Alexa Top-1000 domains for 10 countries). Our findings reveal that the prevalence of web tracking varies across the globe. In addition to location, tracking also seems to depend on factors such as data privacy policies, Internet speed and censorship. We also observe that despite legal efforts for strengthening privacy, such as the EU cookie law, violations are common and very blatant in some cases, highlighting the need for more effective tools and frameworks for compliance monitoring and enforcement.

Keywords: geolocation, internet, privacy, tracking, web

1. Introduction

Third-party web tracking based on user-behavioral profiling has become a major enabling technique for online targeted ads (for business impacts see e.g., [1]; see also Mayer and Mitchell [2] for a discussion on economics and tracking). Tracking is generally performed using cookies, scripts and browser/traffic fingerprinting (see e.g., [3, 2]). Beyond ads/analytics, tracking can also be effectively exploited by government surveillance programs [4]. Indeed, the US NSA has reportedly used Google cookies for targeted hacking/surveillance [5, 6].

Several past studies explore the extent of tracking, evolving techniques used for tracking, and privacy/business implications of tracking. The literature on tracking is rich and becoming very useful to researchers and regulatory bodies. Englehardt and Narayanan [3] recently measured the extent of third-party trackers on Alexa Top-1M websites using the OpenWPM framework [3]. They run their crawler from an Amazon EC2 instance. Fruchter et al. [7] performed another study, albeit at a much smaller scale (Alexa Top-250 country-specific websites), to uncover variations in tracking in four geographical locations (US, Germany, Australia and Japan) of varying policies/laws/cultures. They also used Amazon EC2 machines from different locations. Falahrastegar et al. [8] studied web tracking using Alexa Top-500 country-specific websites for seven countries (USA, UK,

*Corresponding author

Email address: n_samara@ciise.concordia.ca (Nayanamana Samarasinghe)

Australia, China and Egypt, Iran, and Syria) from a single location in the UK. To evaluate the possibility of surveillance via (third-party) cookies and (first-party) plaintext user-identifiers, Englehardt et al. [4] used Amazon EC2 instances from three geolocations: US (Northern Virginia), Ireland (Dublin), Japan (Tokyo).

Although the study by Fruchter et al. [7] indicates that there are significant differences between countries (four in the study), all past studies lack a global perspective, in terms of the number of locations used to measure tracking (1 to 4 countries). Also, all studies were conducted from institutional machines and known IP ranges (university/Amazon), although tracking primarily targets home users (residential machines). Institutional/data center proxies are also prone to be blocked or challenged with CAPTCHAs, to mitigate potential abuse (see e.g., [9]). Therefore, for tracking measurements, the use of residential machines appear to be more appropriate. In some other security-related studies, such as censorship [10], end-to-end connectivity violation [11], and DNSSEC infrastructure management [12], geolocation and/or the use of residential machines have been taken into more serious consideration.

We focus on exploring the effects of geographical variations in tracking as experienced by residential users in various parts of the world. Considering differences in political, social, and cultural factors, we choose 56 countries from across the world for crawling a selected set of web sites, using the Luminati HTTP/S proxy service [13].[1] Using OpenWPM, we automatically crawl different types of website URLs (first parties) including the Alexa Top-1000 global sites (home pages), 1000 URLs hosted on the selected Alexa web domains that were shared via Twitter, and Alexa Top-50 country-specific sites (home pages). Subsequently, we extract third-party information of scripts and cookies from the OpenWPM database, and process them using EasyList rules [15] with BlockListParser [16] to perform privacy-related tracking measurements.

[1] Luminati is a commercial network proxy service, providing residential exit nodes in many countries. Recent work by Mi et al. [14] raises serious doubts about how these machines are recruited (e.g., possibility of compromised machines). However, their methodology for characterizing Luminati nodes appears to be unclear—discussed more in Section 2.

Our results show that the prominence of trackers varies significantly between countries – not only in the country-specific sites, but also for global sites. Furthermore, tracker prominence of inner links of a website appears to be higher than its home page. A significant number of third parties place cookies on websites with long validity periods (e.g., >20 years), egregiously violating any reasonable use scenario, and in some cases existing laws/regulations (e.g., the EU cookie law). Although most trackers are global in nature (mostly owned by US companies), top trackers from countries such as China and Russia appear to operate only within the same country.

Contributions.

1. We extend existing tracking measurement studies in three important directions: (1) crawling websites from 56 countries around the world, representing different political, cultural, regulatory, and Internet speed and freedom situations (cf. four countries used in [7]); (2) the use of residential computers via the Luminati proxy service as opposed to institutional/data center machines; and (3) analyzing web content from home pages and inner links of selected Alexa domains (also studied in [3] for a single location). Our methodology provides a more bona fide, global perspective on tracking.
2. We find that for most cases, a tracker's prominence changes significantly with the geographic location, beyond the dynamic nature of current advertisement/tracking ecosystems (which we also measure separately from Montreal, Canada).
3. We also confirm the findings from existing studies and extend them; e.g., similar to Trevisan et al. [17], we also found that the EU cookie law [18] is violated by most tracking companies/sites. Forwarding web requests to local IP addresses through DNS hijacking was reported for Iran [19]; we also observe similar behavior in Saudi Arabia and Uzbekistan in significant numbers.

2. Related Work

Our work provides a more inclusive, global perspective on tracking, by leveraging existing tools and methodologies from several past studies. Here we summarize a few such efforts. Fruchter et al. [7] measure tracking

variations in four countries with different privacy models (as categorized in [20]): (1) *comprehensive*, protecting all digital data (Germany); (2) *sectoral*, protecting certain types of data such as health-care (USA, Japan); (3) *co-regulatory*, similar to (1), but enforcement is done by industry (Australia); and (4) *mixed/no-policy*, no protection for digital privacy (China/Russia, not studied in [7] due to the non-availability of AWS EC2 instances in those countries). They use Alexa Top-250 country-specific sites, and report significant differences in tracking activities between the countries. For example, the number of third-party cookies in news sites are considerably more in the USA, Japan and Australia, compared to Germany. They further conclude that tracking differences in countries may not solely depend on their privacy models, but also on factors such as policy, regulations and culture.

Tracking primarily leverages third-party scripts and cookies, but other advanced/subtle techniques are also used, e.g., evercookies, cookie syncing, and fingerprinting of browser type, canvas/font, web traffic, and WebRTC, AudioContext and battery-level APIs; cf. [21, 2, 22]. In a comprehensive measurement study, Englehardt and Narayanan [3] recently measured the extent of third-party on Alexa Top-1M websites using the OpenWPM framework [3]. They make 15 types of measurements of stateless and stateful tracking techniques. Their results include many important findings: only few third-parties are present in most sites, news sites hosting the most number of trackers, the use of advanced stateless fingerprinting techniques in the wild, and effectiveness of anti-tracking measures (addons and browser features). They also crawl four internal pages of Alexa Top-10K domains; top 20 trackers are found more prominently on the internal pages compared to the home pages.

Tyson et al. [23] analyze the degree of HTTP header manipulation by middleboxes across ASes in different networks and regions around the world, using Hola [24] (a peer-to-peer VPN service operated by Luminati). They report that 25% of the ASes modify HTTP headers, and the level of manipulation depends on the region and AS type: well-connected regions have fewer caching headers than less-connected regions with costly transit. However, the frequent use of cached data from legacy middleboxes can be exploited.

Using Luminati, Chung et al. [11] propose a novel approach to identify end-to-end violations in HTTP, HTTPS and DNS protocols. They observe that web content sent over HTTP is compressed inflight by some ISPs. They identify a vulnerability where HTTP requests from users are recorded at ISP middleboxes, and the same content is fetched later by third party servers. This allows adversaries to monitor HTTP responses, raising privacy implications.

Pearce et al. [10] designed a measurement platform to assess DNS manipulation attempts for imposing Internet censorship, by leveraging OpenDNS resolvers hosted by ISPs and cloud service providers from 151 countries. They reported that DNS manipulation is heterogeneous across countries, domains and DNS resolvers. Several countries such as Iran, Pakistan, China are found to use DNS manipulation for censorship.

Merzdovnik et al. [25] analyze the effectiveness of current anti-tracking privacy tools on more than 100,000 websites from Alexa Top-200K domains; some of these tools are very effective (over 90% success rates) against stateful trackers, and less successful against stateless fingerprinting trackers. They also report that over 60% of the third-party requests didn't use TLS, which makes it possible for adversaries to passively analyze the unencrypted traffic (i.e., third-party requests and responses). They also highlight the danger of over-reliance of a specific third-party tracker being used in a large number of first-party sites (cf. NSA's alleged exploitation of Google cookies [5, 6]).

According to the EU Internet Handbook [18], the use of *profiling/tracking* cookies require explicit user-consent; session cookies and cookies that are required for essential functionality are exempted. Trevisan et al. [17] use 36,197 popular sites from 25 countries (21 EU and 4 non-EU) to measure the compliance of the EU ePrivacy Directive (also known as the EU cookie law). They also use proxy services from eight EU countries to check variations of tracking cookies based on browsing locations (the EU cookie law's enforcement varies across member states). The authors identify cookies in trackable context by comparing them with a public list of web tracker domains.[2] They find 65% of the web sites fail to comply with the cookie law (i.e., a cookie is set before a cookie accept bar is even displayed to the user). They also ob-

[2] `https://better.fyi/trackers/alexa-top-500-news/`

serve that 80% of the third-party cookies last more than a month, and approximately half of those cookies remain valid for more than a year. We find 22% of the cookies remain valid over a year across EU countries (vs. 23% across all 56 countries).

Mayer et al. [2] observe that third-party web tracking is transitioning from a regulatory vacuum to regulatory frameworks, implemented by government organizations (e.g., US FTC, EU ePrivacy Directive, and self-regulatory programs such as Network Advertising Initiative, Interactive Advertising Bureau).

Degeling et al. [26] analyzed GDPR's impact on Top-500 country specific sites in 28 member states in the EU. They found that GDPR made the majority of companies to make adjustments to accommodate the new regulations. Despite, the authors claim, we find that tracking activities have not changed and most cookie consenting libraries are not meeting the requirements of the GDPR.

Schelter et al. [27] performed a large scale analysis of third-party trackers using the *Common Crawl 2012* corpus. The corpus may contain tracking information of residential as well as institutional users. Since third parties are extracted from static embedding of web pages, transient trackers having dynamic content are not considered. In contrast, our study includes mostly residential computers, and the content we collected is not limited to static trackers.

Web services may be divided into categories e.g., culture, religion, news, sports, etc. To measure tracking variation across different categories, Falahrastegar et al. [8] study seven countries from all continents with different languages using 500 most popular country-specific web sites (crawled from a UK location). Their findings show that some of the top trackers are local to the hosting country of the corresponding first-party website (e.g., websites from China and Iran).

Mi et al. [14] use five residential proxy services including Luminati, for illegal/unwanted/malicious nodes in these ecosystems. They claim that Luminati runs many IoT devices although most exit nodes are indeed residential. However, their methodology for detecting IoT devices inside a NAT requires scanning the internal network (local subnet), which is disallowed by Luminati; thus, such device characterization for Luminati seems to be flawed (also confirmed by Luminati). Luminati also informed us that their proxy software is not supported on any IoT devices (available only for desktop and mobile OSes). Also, Luminati software is installed with explicit user consent, in contrast to the claim by Mi et al. [14]—see Section 4.4 for more issues related to ethics.

3. Background on Luminati and OpenWPM

Luminati. Luminati [13] is a commercial HTTP/S proxy service provider that routes traffic through 35 million residential IPs worldwide. The service operates over Hola [24] and applications built using the Luminati Monetization SDK [28]—residential users without a paid subscription. Luminati is gradually transitioning from Hola to the SDK model. However, at the time we ran our experiments, Hola was used comprehensively for Luminati's exit node infrastructure. Routing in Luminati goes as follows: a Luminati client makes a proxy connection to a Luminati proxy server (super proxy); the server forwards the request to an exit node (peer proxy); and the exit node forwards the response to the super proxy, which in turn is sent back to the Luminati client. Luminati enables selecting exit nodes by country (or city/ASN at a higher cost), and allows the same exit node to be used in subsequent requests by using the "sequential session (IP) pool" option. Switching the IP address of an exit node in the pool can be configured based on the number of maximum requests and session duration parameters, or at random. Luminati also allows controlling DNS resolution to happen at the super proxy (Google Public DNS), or the exit node. We choose a sequential pool of pre-established sessions to run a group of requests to target sites. Also, we configure DNS resolution to happen at a super proxy (US), to prevent DNS localization of web site domains at exit nodes so that trackers of the same first-party site are comparable between countries; e.g., when crawling `amazon.com`, we do not want the exit node to retrieve content from a regional first party site e.g., `amazon.com.mx` due to DNS localization. This is unlikely to influence the comparison of regional trackers as their DNS resolution remains unaffected. We verified from our dataset that the domains of regional third parties (e.g., ebay.de, ebay.it, ebay.fr) remained unchanged. Luminati supports super proxy IP caching where three super proxy IPs in the cache are available to service requests, eliminating unnecessary timeouts due to distant super proxies.

The IP address of a user, connecting to a websites through a proxy can be identified using the *X-Forwarded-For* [29] request header. Adding the IP address of the user to *X-Forwarded-For* request header by a proxy defeats the purpose of being anonymous [30] with the connecting server. Luminati has been adding this header to requests in the past [31]. However, we did not find any *X-Forwarded-For* headers that contain the IP address of users in the requests initiated through the Luminati proxy manager.

OpenWPM. OpenWPM [3] is a web privacy measurement framework that automates the crawling of a large number of URLs and reduces engineering efforts for tracking studies. A built-in proxy is available, which we replace with Luminati. We configure OpenWPM (ver: 0.7.0) for stateless crawling (each new page-visit uses a separate browser profile), as we are primarily interested in the location-related aspects of tracking. Instrumentation results are stored in a local SQLite database; we modify the database schemas to record additional information, e.g., the exit node's IP address, AS details, and location (country). We launch three browser sessions simultaneously through OpenWPM; we could not further increase the number of parallel sessions due to performance issues which would crash the crawling sessions (system configuration: AMD FX8350, 8GB RAM, Ubuntu 16.04, Gigabit Internet).

4. Methodology

We use the Luminati proxy manager [13] to run experiments from 56 countries, and the OpenWPM privacy measurement framework [3] for automating browser data collection and tracker analysis on a selected list of URLs. In this section, we expand on our country/URL selection, and define trackers and their prominence (largely based on [3]); see Fig. 1 for an overview of our experimental setup (for our Luminati, OpenWPM configuration, see Section 3).

The requests from OpenWPM crawls are proxied via Luminati, so that they go through exit nodes in the country of our choice. Luminati passes the response from exit nodes back to OpenWPM, which processes the response data, extracts privacy related measurements, and stores them in a database. We then query the database to analyze the measurement data and compute various metrics.

Through Luminati, we process a total of 68,800 URLs. These sites include, Alexa Top-1000 (global) and Top-50 (country specific) URLs from 56 countries, and 1000 URLs shared via Twitter for 10 selected countries. Each URL request takes 1.16MB of bandwidth on average (including repeated attempts for failed/timed-out requests). We run the experiments between June 1 and July 8, 2017. Using Luminati is expensive. [3]

We only considered successful third party requests for our analysis. Those URL responses with client errors (4xx status code) and server errors (5xx status code) are eliminated from the analysis. Such failures are attributed to many reasons (e.g., authentication issues, timeouts, censorship).

4.1. Country and first-party site selection

The use of residential machines from all countries/regions/cities would be ideal for our goal. However, using Luminati is costly, and it also lacks exit nodes in certain countries (e.g., North Korea). Covering several regions with various political and socio-economic situations, we select 56 countries. We list in Table 1 the countries in various regions used in our experiments. Our selection is influenced by Freedom House [32], and Swire and Ahmad [20].

The 2050 distinct URLs that we use for crawling include: (1) home pages of Alexa Top-1000 global domains; (2) 1000 popular URLs that are shared via Twitter from the Alexa Top-1000 domains, excluding home pages, and links to media (e.g., images, audio and video) and text files (which may not host any tracker); and (3) home pages of Alexa Top-50 country-specific domains.

We extract Twitter URLs using Tweepy [33] that internally uses the Twitter streaming APIs to access the global stream of Twitter data. Twitter mandates that a client filter the streamed data according to a specific criterion. To not omit streams from any parts of the world, we use: `twitterStream.filter(locations = [-180,-90,180,90])`. Assuming geotagging is turned

[3] With the cheapest Luminati *Starter residential* package, it costs USD 12.50/1GB (for 40GB, with a minimum monthly commitment of USD 500). Hence, we incurred USD 14.50 for 1000 URLs. Therefore, for 56 countries, it will cost USD 812 to process 1000 URLs. For 1 million URLs (cf. [3]) from 56 countries, the cost will be USD 812,000. We thus had to limit the number of crawled URLs.

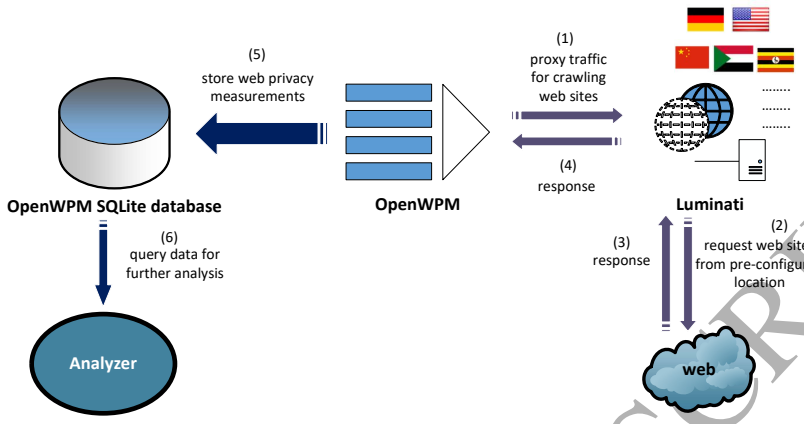

Figure 1: Our system setup.

Asia-Pacific	Australia (AU), Bangladesh (BD), China (CN), India (IN), Japan (JP), Malaysia (MY), Myanmar (MM), Pakistan (PK), Philippines (PH), Singapore (SG), South Korea (KR), Sri Lanka (LK), Vietnam (VN)
Americas	Argentina (AR), Brazil (BR), Canada (CA), Colombia (CO), Cuba (CU), Ecuador (EC), Mexico (MX), United Sates (US), Venezuela (VE)
Europe	Estonia (EE), France (FR), Germany (DE), Great Britain (GB), Hungary (HU), Iceland (IS), Italy (IT), Turkey (TR), Switzerland (CH)
Eurasia	Armenia (AM), Georgia (GE), Kazakhstan (KZ), Russia (RU), Ukraine (UA), Uzbekistan (UZ)
Middle East and North Africa	Bahrain (BH), Egypt (EG), Iran (IR), Israel (IL), Jordon (JO), Lebanon (LB), Libya (LY), Morocco (MA), Saudi Arabia (SA), Tunisia (TN), United Arab Emirates (AE)
Sub-Saharan Africa	Ethiopia (ET), Kenya (KE), Nigeria (NG), Rwanda (RW), South Africa (ZA), Sudan (SD), Uganda (UG), Zimbabwe (ZW)

Table 1: List of regions and countries.

on, this filter selects tweets from all around the world using the *locations* filter. We select the most shared URLs from the Alexa Top-1000 domains.

4.2. Tracker identification and prominence

We define the third parties as follows. (1) Third-party scripts: the domain on which the third-party script runs is different from the domain of the first-party site. (2) Cookies: the cookie's domain is different from the domain of the first-party site.

Not all identified third parties may necessarily be trackers. Third party domains can be trackers, advertisers, or simply content embedded on a first-party site. We use BlockListParser [16] to filter third parties in a tracking context with the aid of a set of ad-blocking filtering lists as used by the AdBlock browser extension: *EasyList* and *EasyPrivacy* [15]. *EasyList* tracking protection lists con-

tain rules to identify trackers which are also advertisers, while *EasyPrivacy* identifies non-advertising trackers [3]. This filtering is in line with previous studies; cf. [3, 7]. For analysis, we keep trackers that exist on at least two first-party sites (similar to [3]). Since advertisers in certain circumstances can play a dual role as trackers, we emphasize that the identified trackers in our analysis may fall into a lower bound of trackers in reality; more sophisticated filtering is difficult (e.g., some third parties directly, or through their parent organizations, may act as genuine content providers [34]). For example, Google receives a large proportion of content related third-party requests that do not fall into the categories of tracking or advertisements.

To identify tracker domains based on third-party scripts or cookies, we use *public suffix + 1* (*PS+1*) of the script URL or the cookie domain (along with Mozilla's Public Suffix List[4] as in [3]). For example, if the script URL is `http://tpc.googlesyndication.com/sodar/d5qAyLYU.js`, then PS+1 of the domain is googlesyndication.com; if the script is included as a dependency in `http://oneindia.com`, then googlesyndication.com is a third-party tracker.

Tracker prominence. A possible limitation of measuring a tracker's activity using the number of first parties on which the tracker is present is that the tracker may have a low first-party count, when a limited number of first-party sites are used. To properly rank trackers' prominence, we use the following metric from Englehardt and Narayanan [3]: $Prominence(t) = \sum_{edge(s,t)=1} \frac{1}{rank(s)}$; $edge(s,t)$ indicates third-party t's presence on site s. This metric mitigates the distortion of a tracker's importance due to the selection of a small set of first-party sites (as in our case, 1050- 2050 URLs per country). Such a small set of first-party sites may not include all first parties where a particular third-party is highly prevalent.

Comparing countries. To compare the extent of tracking between countries, we treat the prominence values of trackers of each country as a group, and we compute non-parametric Kruskal-Wallis (KW) rank averages (assuming groups are independent). Countries with a higher rank average should have a higher level of tracking and vice-versa. Furthermore, the rank averages of all the countries can be used to perform the *KW test* to determine if the level of tracking between countries is independent of each other or not. In a KW test, a null hypothesis is initially assumed where all samples (i.e., groups) come from identical populations. If the KW test value is greater than the critical chi-square value, the null hypothesis is rejected, proving at least one group comes from a different population. A similar approach was adopted by Fruchter et al. [7] for comparing tracking activities between four countries.

4.3. Dynamicity of trackers

Since ad exchanges leverage a Real-time Bidding (RTB) auction based model where only winning bidders are allowed to serve content to users [35, 36], web trackers are also expected to be dynamic in nature. However, dynamicity of trackers have not been discussed in previous large scale measurement studies (e.g., [3, 7, 8]). To establish ground-truth on the limits of dynamic behaviors of trackers, we conducted several experiments with Alexa Top-1000 sites. We calculated the difference of the number of first parties for each tracker as observed from two different ISPs within Montreal, Canada; we performed 12 tests simultaneously from both ISPs, and at different times of the day, over a period of two months, where each test took approximately 4 hours to complete.

We use *z-score*[5] to assess the variation of trackers; z-score measures the number of standard deviations of the signed distance between a data point and the mean of a distribution.[6] If the data point is greater than the mean, the z-score is positive, otherwise it is negative. Overall, z-scores for our observations lie between -0.4 and +0.4. For simultaneous runs from both ISPs, the differences and z-score values of the number of first parties for the Top-5 trackers are: advertising.com (223, 0.36), pubmatic.com (192, 0.27), adsafeprotected.com (140, 0.11), moatads.com (75, -0.09), scorecardresearch.com (68, -0.11). Similarly, when measured from a particular ISP at different times, the differences and z-score values for the

[4] Hosted at: `https://publicsuffix.org`; a public suffix is defined as "one under which Internet users can (or historically could) directly register names."

[5] `https://en.wikipedia.org/wiki/Standard_score`

[6] Unlike standard deviation, z-score is used to compare scores from different distributions [37]. Also, z-score determines whether a given value is typical in a data set.

Top-5 trackers are: openx.net (217, 0.39), googlesyndication.com (114, 0.05), adnxs.com (109, 0.03), gstatic.com (73, -0.09), yandex.ru (73, 0.09). These values change when measured at different times, although the z-scores always remain within -0.4 and +0.4. We also computed the Pearson correlation coefficient for the number of first parties that trackers are found in different runs (different ISPs, and at different times of the day); our Pearson coefficient turned out to have a highly positive linear correlation (0.9), implying that the trackers identified in independent runs follow a strong linear relationship. Therefore, for the overall tracking ecosystem, the dynamicity of trackers does not appear to have adversely impacted the interpretation of results of our measurement study.

4.4. Ethical issues

We access residential users' Internet connection through Luminati, which is a paid service. We do not compromise the security or privacy of users (of exit nodes) beyond using their internet connection, which they have agreed to when signing up with Luminati. These users include those using Hola [24] clients and applications built leveraging the Luminati monetizing SDK [28] without a subscription. Hola and Luminati explicitly mention the sharing of internet connection to their users. Furthermore, we do not store the response content returned by the websites, except the measurements for trackers.

Some websites that we crawl (Alexa top sites and Twitter-shared URLs) may be censored in a few countries. Other than Egypt,[7] we are unaware of any place where attempts to access blocked/censored content will trigger legal problems for a user. During our tests, the new law in Egypt that threatens imprisoning those browsing censored web sites did not exist. Besides, the sites crawled from Egypt are not subjected to censorship according to the Citizen Lab dataset [38]. We are unable to get the consent from targeted users owning Luminati exit nodes, as we do not have their contact information. However, we reached out to the internal Research Ethics Unit of our University, and explained our experiments; they did not object to our methodology and did not require us to go through a full ethics evaluation.

4.5. Limitations

We use Luminati's residential exit nodes for measuring web tracking from a home user's perspective. However, we have no control over such nodes (compared to using more reliable university/EC2 infrastructures). Here we list some issues that may affect our results. (1) Web tracking may depend on the browsing history of a specific client as identified by its IP address. Thus our results may be influenced by the browsing history of the Luminati exit nodes, which is beyond our control. This is an inherent limitation of using residential IPs as opposed to university/Amazon IPs. However, our connections are not effected by local cookies or other browsing data (only share the same IP address). (2) We crawl websites via OpenWPM in a sequential order over the period of five weeks. Hence, time dependent trackers (if any) may affect our results. Furthermore, the number of trackers on first party sites may grow or shrink with time. This is due to many reasons, including technological advancements of tracking techniques [39], outages, performance issues with tracking services, and ISP filtering [40, 41]. A comprehensive study of such dynamic behaviors and uncertainties of tracking at a global scale is beyond our scope as it requires repeated tests, which is not pragmatic due to the high cost of using Luminati. However, we measured dynamicity of trackers via two ISPs from two locations, and found the impact to be limited to our measurement criteria (see Section 4.3). (3) Tracking context of some Google trackers (e.g., google.com, gstatic.com, youtube.com) are omitted from our work as they are not proxied by Luminati. We realized this limitation during our experiments.[8] However, most Google-owned tracking domains remain unaffected, i.e., proxied through Luminati. We manually verified this for all top tracker domains in our list. (4) The EasyList [15] filter that we use to identify third party domains participating in a tracking context may not have adequate coverage in all countries, although it can filter most trackers from international web pages. Therefore, our results may not include trackers that are not identified

[7] News article (Aug. 19, 2018): `https://www.telegraph.co.uk/news/2018/08/19/egyptians-face-jail-accessing-banned-websites/`

[8] Luminati prohibits "Any form of outbound automated Google search queries", but mentions no other Google related restrictions.

by EasyList rules. However EasyList offers several supplementary filter lists [42] to support several non-English domains (e.g., German, Italian, Dutch, French, Chinese, Bulgarian, Arabic, Czech, Slovak, Lithuanian and Hebrew). The coverage of these supplementary lists are still unknown. (5) If a first party website intentionally uses a third party for tracking the site visitors, we do not distinguish such trackers from other third parties participating in a tracking context.

5. Trackers vs. geolocation

In this section, we explain the analysis process followed by the results. Unless otherwise stated, the tracking context is measured using third-party scripts on the Alexa Top-1000 global domains.

We first check the presence of top-10 trackers in Alexa Top-1000 domains in all countries. For brevity, we highlight the results from 15 countries with most significant differences across regions; see Fig. 2. The top-10 trackers are determined based on the average percentage of first-party sites across 15 countries. If multiple instances of the same tracker are found on a particular first-party site (e.g., several scripts from a single tracker domain), we count them separately.

In Fig. 2, darker-shade trackers have more presence in first-party sites compared to lighter shade ones. We calculate the tracker percentages for each country based on the first party count for the specific tracker over the total first-party count of the specific country. Highest percentages for googlesyndication (25.6%) and doubleclick (23.2%) trackers are observed in Russia and Ethiopia respectively; the percentages are relative to other trackers observed from the same country. These two trackers are also prominent in all other countries. In contrast, some trackers are not seen in certain countries (blank cells in Fig. 2).

China and Iran have a relatively low percentage of trackers. Google advertisements are sanctioned in Iran by United States Office of Foreign Assets Control (OFAC) [43]. Schelter et al. [27] observed a similar pattern in their study, and they justify this behavior due to political factors including lack of democracy and freedom of the press.

We also identify the top-10 tracking organizations; we use *pywhois* [44] to locate organizations from corresponding domains; see Fig. 3. Google has a clear domination across the world. Note that despite Google services, e.g., Google Search, Maps, Docs, Mail being censored in China [45], Google trackers remain active in China on uncensored websites.

6. Overall tracker prominence

In this section, we analyze the differences in tracking, using prominence and KW rank metrics (Section 4.2), and compare 56 countries; see Fig. 4. UK and Armenia have the highest prominence values, while Iran and Ethiopia have the least. The latter countries are known to have less media/Internet freedom [19, 32]. But countries such as Morocco, Singapore, Venezuela, Mexico and Rwanda have relatively higher prominence values although they rank low in media/Internet freedom, showing the presence of other factors influencing tracking. We discuss the impact of few of those factors such as Internet speed, censorship and browser user agents in Section 8.

We summarize prevalence of top trackers (in terms of the average of raw count in countries) in different regions in Fig. 5. In general, Europe has the highest count compared to others, despite the EU cookie law. Degeling et al. [26] claims, GDPR didn't have a noticeable change in tracking although it made the web more transparent by having the website owners updating their privacy policies.

Our results from the KW test show that the tracker prominence among different countries are independent of each other ($x^2 = 83.64, df = 55, p = 0.05$). This is because the null hypothesis of the KW test is rejected as the KW test value ($x^2 = 83.64$) is greater than the critical chi-square value (73.311) [46] with 55 degrees of freedom (df), where p-value (used to accept/reject the null hypothesis) is 0.05. Therefore, the prominence of trackers varies with different browsing locations that are independent of each other.

Comparing prominence between home pages and Twitter URLs. We also compare countries based on tracker prominence in Alexa Top-1000 home pages and Twitter-shared URLs. For this experiment, we consider 10 countries across all regions. We calculate the prominence values of trackers in each country for the home pages and Twitter URLs; see Fig. 6.

It is apparent that the prominence values of trackers in Twitter URLs are significantly higher compared to home

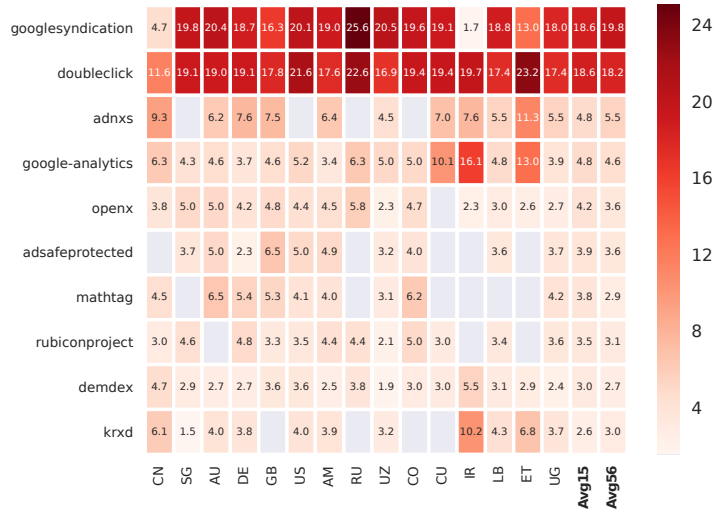

Figure 2: First-party percentages for script-based trackers across 15 countries; *Avg15* and *Avg56* represent average percentages for the selected 15 countries and all 56 countries, respectively.

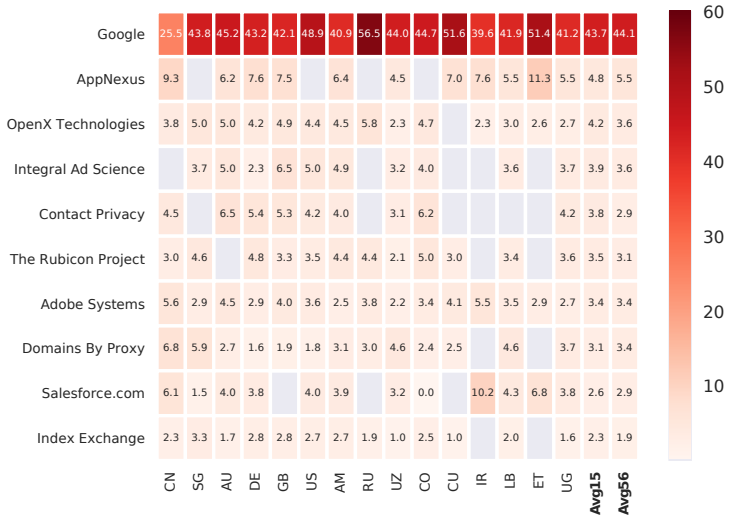

Figure 3: First-party percentages for tracking org.

pages in *all* selected countries. Englehardt et al. [3] noticed a 6%–57% increase of third-party presence on first parties (for top 20 third parties) with inner URLs as opposed to their home pages. In our experiments, the increase of prominence in Twitter URLs is between 7%–28%. However, we take into account the prominence of all third parties available. Therefore, it appears that increase of third party presence in inner URLs is relatively higher for the top trackers.

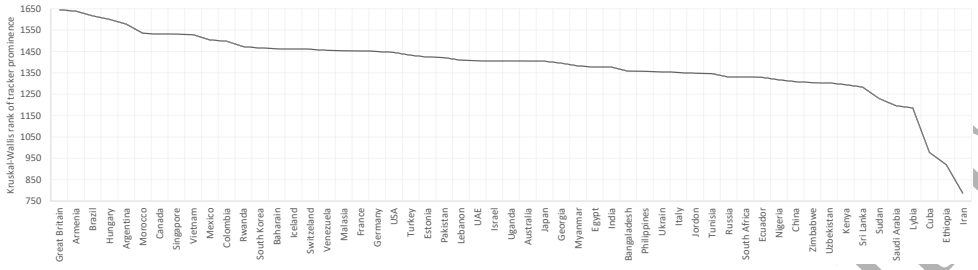

Figure 4: Prominence of tracker scripts across 56 countries.

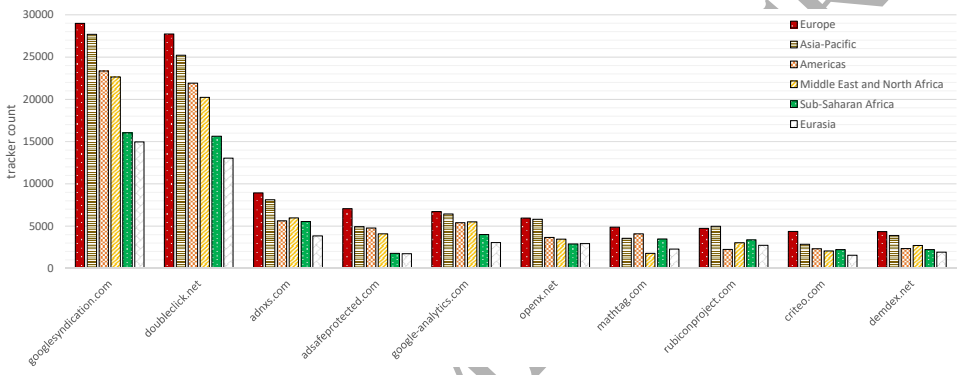

Figure 5: Average of top trackers in different regions.

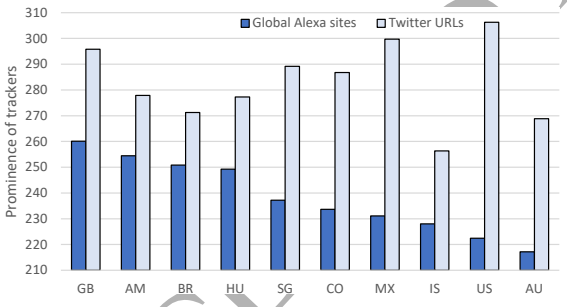

Figure 6: Prominence of tracker scripts: Alexa global sites vs. Twitter URLs.

Trackers in country-specific sites. Most trackers across the world are hosted from US domains. However, similar to the observations in Falahrastegar et al. [8], we note an exception from Top-50 country-specific first-party sites in China and Russia, where the top trackers for both third-party scripts and cookies originate from the same country. In China, baidu.com tops the first-party count in both tracking scripts (93) and cookies (5). Similarly, in Russia, yandex.ru is a leading tracker having the highest first-party count for tracking scripts (427) and cookies (25). The difference in approaches between Falahrastegar et al. and ours is that in the former, 500 country-specific first party sites are used (from the same location), while we use Top-50 country specific sites (from 56 countries); they report more baidu.com count (approx. 2000), although they do not clarify the tracking context.

7. Cookie validity durations

Similar to Trevisan et al. [17], we also found that the EU cookie law is not complied by most tracking companies in the EU and non-EU countries; see Fig. 7. Many cookies have a validity period over 20 years, and some up to 7988 years (e.g., rubiconproject.com, rfihub.com). Overall, UK and US have the highest counts of these cookies, while Iran, Cuba, Ethiopia and Libya have the

least. We did not use the data collected from Kazakhstan for our analysis, as it is impacted due to slow connections at the respective exit nodes.

Access country	>1year	>20 years
Great Britain	10,516	3618
Germany	5047	1956
Hungary	5071	1866
Italy	5000	1853
France	4250	1801
Estonia	2692	1267

Table 2: Number of tracking cookies with validity periods (EU).

Trevisan et al. [17] found 65% of the websites in 25 countries (21 from EU) install tracking cookies. In comparison, we found that, for the nine European countries, 60% of first-party sites set tracking cookies without consent, which is even higher than our global average (56.2% sites in the 56 countries; see Degeling et al. [26] for technical issues in GDPR compliance and common cookie consent implementations).

In addition to third party cookies, some first party cookies (e.g., doubleclick.net, paypal.com) contain unique pseudonymous identifiers, although they do not include Personally Identifiable Information (PII) [47]. We did not find attributes in first party cookies containing any identifiable PII

The top-5 domains of tracking cookies with over a year validity are as follows (the number of cookies, first-party percentages): scorecardresearch.com (23,171, 0.015%); rubiconproject.com (12,680, 0.008%); rfihub.com (12,105, 0.008%); advertising.com (11,042, 0.007%); and adtechus.com (9940, 0.006%). We also checked their privacy policies (Sept. 2, 2018). They do not mention their cookie validity periods, but claim to be in compliance with EU privacy laws (including GDPR [48]). The opt-out mechanism of scorecardresearch.com is also cookie based [49], i.e., opt-out is not possible when cookies are blocked or deleted. Top-6 EU countries with the most number of cookies with long validity periods are listed in Table 2. Top-10 EU specific domains with the highest number of tracking cookies are listed in Table 3; for each domain, we also list the countries from which the requests are originated, and the country where the third-party domain is registered (most registered in France and Germany).

8. Factors other than geolocation

8.1. Internet speed

Tracking appears to vary proportionally with Internet speed in a country; see Fig. 8 (we use Akamai's report [50] on global Internet speed as of June 1, 2017). The countries right to the vertical dotted line in Fig. 8 have a higher tracker prominence (marked red ▾). Tracker prominence values are high for Armenia, Colombia, Lebanon and Uganda, although they have relatively low Internet speed. Resource intensive tracking sources included in first-party sites may not completely load with a slow connection, increasing the rate of request failures.[9]

The failure rates of OpenWPM requests (following its re-connection attempts) vary between 11.21% – 19.04%. The highest failure rates are in United States (19.04%), Great Britain (18.26%) and Germany (18.13%). These countries still rank high in tracking prominence. We also checked the connection failures more closely, and observed that such failures are more common for trackers than the first-party sites.

To understand the impact of HTTP/S errors on tracker prominence, we calculated the difference of KW ranks of all HTTP/S requests vs. requests without client/server errors. Fig. 9 shows the KW rank of requests with errors. Although the tracker prominence and the rate of failures are not proportional in all 56 countries, the KW rank of request errors are high in Great Britain (686), United States (673) and Mexico (576), while they are lower in Uzbekistan (186) and Ethiopia (141).

8.2. Censorship

Apparently, there is a direct relationship between Internet/media freedom and tracking prominence—more open countries seem to attract more trackers; see Fig 10 (for

[9]The average webpage size is growing significantly, every year; in 2017, it is approximately 2.5MB, part of which is attributed to trackers, see e.g., KeyCDN (`https://www.keycdn.com/support/the-growth-of-web-page-size/`). For example, CNN's home page size is 4.7MB and the page creates 349 HTTP requests (as of Sept. 25, 2018; tested using `tools.pingdom.com`).

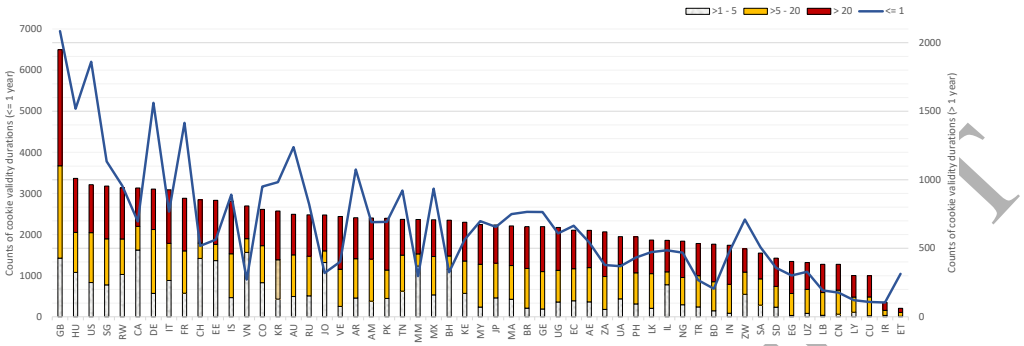

Figure 7: The number of cookies vs. validity period: counts of cookie validity periods $\leqslant$ 1 year are shown by the line (left y-axis), and the rest (> 1 year) are shown with bars (right y-axis). All these cookies are set without user consent.

Tracker domain	Access country	Reg.	Count
smartadserver.com	DE, EE, FR, GB, HU, IT	FR	4005
angsrvr.com	DE, FR, GB, HU, IT	DE	1570
criteo.com	EE, FR, GB, HU, IT	FR	1461
ml314.com	DE, EE, FR, GB, HU	FR	920
theadex.com	DE, EE, GB, IT	FR	665
yieldlab.net	DE, GE, HU	DE	420
visualdna.com	DE, GB, IT	GB	417
semasio.net	DE, GB	DE	392
switchadhub.com	EE, FR, GB, HU, IT	GB	303
ligadx.com	HU	DE	280

Table 3: Domains of top-10 tracking cookies registered in EU countries.

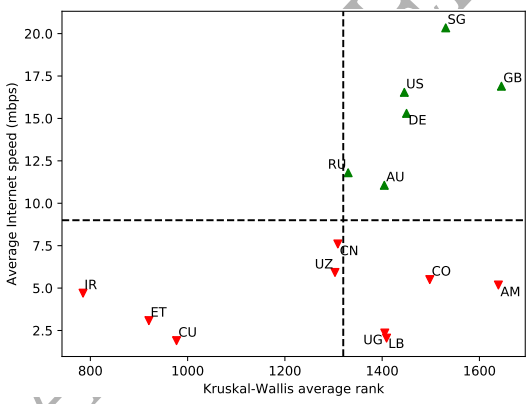

Figure 8: Internet speed vs. tracker prominence.

clarity, we show only 15 countries, but a similar trend is observed for all 56 countries). We divide the countries into three categories based on the 2017 Freedom of the Press rankings [32]; countries marked in red (▾) are considered to be *free*, amber (■) ones *partially-free*, and green (▴) ones *not-free*. All the not-free countries have a lower tracker prominence. Note that, although Ethiopia shows a higher percentage of trackers in Fig 3, those values are relative to the country.

We also analyzed HTTP response codes. While there were many codes other than 200 (OK), those with 403 (Forbidden Host) are interesting: majority of third parties included in these first party sites appear to be hosted on local IP addresses (e.g., 10.10.34.34, 192.168.1.1). Other studies [19, 10] also reported similar behavior in Iran as ours (87 occurrences), where DNS hijacking is used for censorship; a blocked site is redirected to a web page running on a local IP address that is accessible within Iran. In addition to Iran [19], we also observe the same behavior

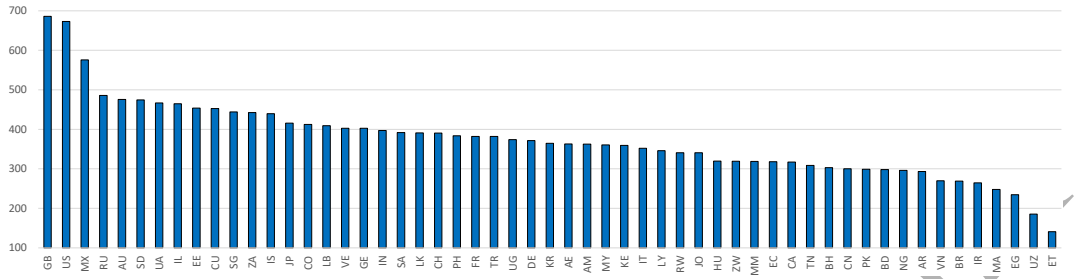

Figure 9: KW Ranks highlighting errors of HTTP/S requests.

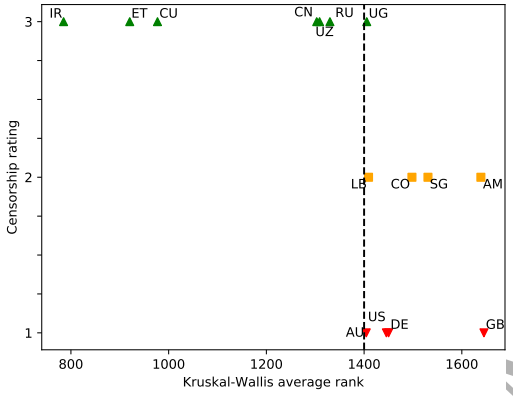

Figure 10: Censorship rating vs. tracker prominence.

with SA (173), UZ (46), NG (28), GB (22), PK (19), TN (11), and US (9). Note that the 403 response code is also returned when appropriate authorization is not provided (e.g., a non-public page).

8.3. Browser user-agents vs. tracking

A user-agent, as sent with an HTTP request, can help identify a user's device, browser/OS versions, and even a specific user (although not very accurately) [51]. Currently, OpenWPM supports only Firefox. We modify OpenWPM with a list of user agents[10] supporting different browser/platform types. Considering four popular browsers—Chrome, Firefox, IE and Safari, we use a total of forty user agents with different desktop OSes (Windows, Mac OS X, and OpenBSD); a random user-agent is picked for each crawl. This allows an unbiased approach in simulating requests made from different browsers (instead of sending a series of requests with the same user-agent). We run the tests for each browser type at a time (i.e., each browser type is tested equally).

We summarize prevalence of top trackers (scripts vs. cookies) for common browser user-agents; see Figs. 11 and 12. Some trackers appear significantly more than the rest across all user agents for Chrome, Firefox, IE and Safari—e.g., googlesyndication.com and doubleclick.net in tracking scripts, and adnxs.com and rubiconproject.com in tracking cookies. Surprisingly, some trackers do not appear at all for certain browser types. We validated such unusual cases with manual inspection using *Chrome DevTools* [11], and similar tools in other browsers (e.g., F12 Developer tools [12]).

9. Data protection laws vs. tracking

We summarize below data protection laws in different regions and explain their relevance to tracker prominence (see Section 6), based on DLA Piper [52]. Overall, countries with higher tracker prominence also have relatively tougher data privacy regulations, implying whether such regulations are properly enforced.

Asia Pacific. No specific laws or regulations exist relating to data privacy except in South Korea (prominence score:

[10] Extracted from: `http://www.useragentstring.com/pages/useragentstring.php`

[11] `https://developers.google.com/web/tools/chrome-devtools/`

[12] `https://docs.microsoft.com/en-ca/microsoft-edge/devtools-guide`

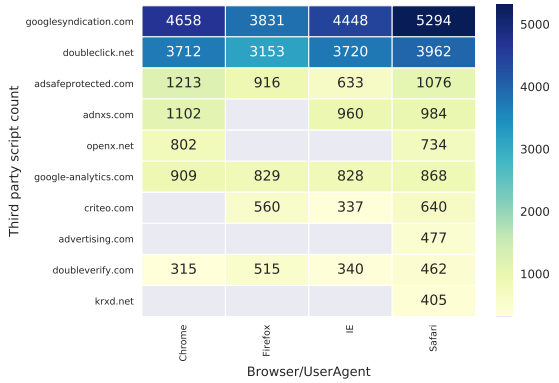

Figure 11: Tracking scripts vs. user-agents.

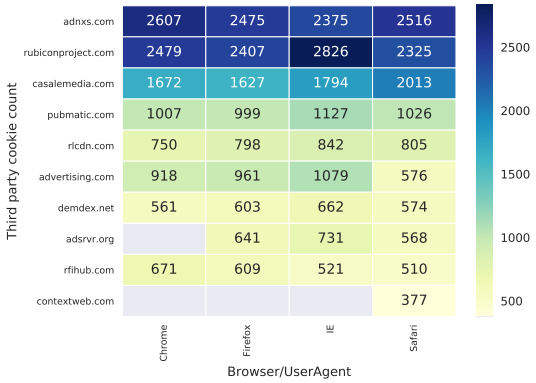

Figure 12: Tracking cookies vs. user-agents.

1466), with a fairly higher tracker prominence. In South Korea, cookie, log and IP information usage is governed by *IT Network Act*, and requires to get opt-out consent from users. Location information of users is regulated by the *LBS Act*. Australia (1404) leverages its *Privacy Act*, state and privacy laws to regulate e-privacy and the collection of location data to some extent.

Americas. Canada (1531), United States (1446) and Mexico (1504) have a higher tracker prominence. On top of provincial laws, Canadian *Personal Information Protection and Electronic Document Act (PIPEDA)* applies to consumer and employee personal information. In the US, *Federal Trade Commission (FTC)* ensures businesses take reasonable minimal data security measures to ensure consumer privacy. In contrast, South American countries (only Argentina and Uruguay are covered by DLA Piper) lack privacy laws, which also have relatively higher tracker prominence values.

Europe. EU's General Data Protection Regulation (GDPR) [48] is in effect since May 2018, governing all its member states alike. While the existing e-privacy directives in EU are complied by its member states, there is no clear indication of any reduction in tracking activities due to these regulations. Although, France, Switzerland and Italy are more strict in applying e-privacy laws compared to Germany, our results indicate France (1453) and Switzerland (1460) have comparable tracker prominence as Germany (1450). Switzerland requires explicit consent from users before data is collected, and personal data (e.g., stored in cookies) is deemed to be sensitive. France requires traffic data to be anonymized or erased, and not use location data without explicit consent. In Italy (1352), traffic data is supposed to be removed when no longer required, and cannot be held for more than 6 months. According to UK's (1645) Privacy and Electronic Communications (PEC) act, traffic data needs to be erased when not required and can be used with consent for value added services; nevertheless, UK has the highest tracker prominence.

Eurasia. Russia (1330) and Ukraine (1355) do not have specific privacy legislations, but their tracker prominence values are lower than most countries.

Middle-East. Baharain's (1463) tracker prominence is relatively high and it lacks any privacy laws. Saudi Arabia (1196) also has no privacy rules. In UAE (1355), although its penal code does not provision regulations for Internet privacy, the general laws contained therein can be applied for online privacy. Both Saudi Arabia and UAE have lower tracker prominence values. Tracker prominence in Egypt (1377) is also low; its 2014 constitution provides clear guidelines on Internet security, but not about privacy. In 2017, the Egyptian government cracked down on encryption and circumvention tools [53].

Sub-Saharan Africa. Most countries in this region have a lower tracker prominence. South Africa (1330) doesn't appear to have laws to regulate privacy. However, Nigeria (1317) has regulations for electronic communi-

cation/privacy rights with respect to cookies and location data.

10. Conclusions and Future Direction

We observe a significant variation of trackers on first-party sites between countries. Some Google trackers (e.g., doubleclick and googlesyndication) on average have an extensive presence compared to other trackers (cf. [54]). The UK and Armenia have the highest tracking prominence, while Ethiopia and Iran have the least. We observe a significant number of cookies valid for many years (>20) in EU countries and elsewhere.

This study can be extended to compare tracking between residential IPs and cloud IPs in multiple countries.

Several other factors also influence tracking beyond location. The countries that enjoy a greater freedom of expression and information flow show a stronger presence of trackers. We also noticed several third-party requests are censored in Iran and few other countries. Also, in general, having stronger privacy regulations does not limit tracking in any significant way.

Tracking is not limited to a particular desktop or a device, and can cross the boundary between multiple devices [55]. Effects of geolocation in cross device tracking could be an interesting future direction. As Luminati does not proxy some Google domains, future work should also consider similar alternative services, if available.

There are many other forms of tracking in addition to third party scripts and cookies (e.g., browser fingerprinting, Flash cookies) which are excluded in our work. The impact of such tracking techniques with geolocation can be studied as future work.

Currently, OpenWPM supports only the Firefox browser. Instead of our user-agent manipulation, measuring with other real browsers may provide a more comprehensive view (but may require significant engineering effort). For example, we could then easily compare tracking prevalence between the Tor and other browsers. Future work may also use the OpenWPM WebExtension tool [56] for privacy measurements in a cross-browser environment using the WebExtension API (supported by all common browsers).

References

References

[1] Recode.net, Google leads the world in digital and mobile ad revenue, News article (July 24, 2017). https://www.recode.net/2017/7/24/16020330/google-digital-mobile-ad-revenue-world-leader-facebook-growth.

[2] J. R. Mayer, J. C. Mitchel, Third-party web tracking: Policy and technology, in: IEEE S&P'12, San Francisco, CA, USA, 2012.

[3] S. Englehardt, A. Narayanan, Online tracking: A 1-million-site measurement and analysis, in: CCS'16, Vienna, Austria, 2016.

[4] S. Englehardt, D. Reisman, C. Eubank, P. Zimmerman, J. Mayer, A. Narayanan, E. W. Felten, Cookies that give you away: The surveillance implications of web tracking, in: WWW'15, Florence, Italy, 2015.

[5] WashingtonPost.com, NSA uses Google cookies to pinpoint targets for hacking, News article (Dec 10, 2013). https://www.washingtonpost.com/news/the-switch/wp/2013/12/10/nsa-uses-google-cookies-to-pinpoint-targets-for-hacking/.

[6] E. Felten, J. Mayer, How the NSA piggy-backs on third-party trackers, Blog article (Dec 13, 2013). http://www.slate.com/blogs/future_tense/2013/12/13/nsa_surveillance_and_third_party_trackers_how_cookies_help_government_spies.html.

[7] N. Fruchter, H. Miao, S. Stevenson, R. Balebako, Variations in tracking in relation to geographic location, in: WTSP'15, San Jose, CA, USA, 2015.

[8] M. Falahrastegar, H. Haddadi, S. Uhlig, R. Mortier, The rise of panopticons: Examining region-specific third-party web tracking, in: TMA'14, London UK, 2014.

[9] Ghost Proxies, The difference between residential and datacenter proxies, Blog article (2019). http://ghostproxies.com/blog/2016/06/residential-datacenter/.

[10] P. Pearce, B. Jones, F. Li, R. Ensafi, N. Feamster, N. Weaver, V. Paxson, Global measurement of DNS manipulation, in: USENIX Security Symposium'17, Vancouver, British Colombia, Canada, 2017.

[11] T. Chung, D. Choffnes, A. Mislove, Tunneling for transparency: A large-scale analysis of end-to-end violations in the Internet, in: IMC'16, Santa Monica, CA, USA, 2016.

[12] T. Chung, R. van Rijswijk-Deij, B. Chandrasekaran, D. Choffnes, D. Levin, B. M. Maggs, A. Mislove, C. Wilson, A longitudinal, end-to-end view of the DNSSEC ecosystem, in: USENIX Security Symposium'17, Vancouver, British Colombia, Canada, 2017.

[13] Luminati, Luminati proxy network, `http://luminati.io/`.

[14] X. Mi, Y. Liu, X. Feng, X. Liao, B. Liu, X. Wang, F. Qian, Z. Li, S. Alrwais, L. Sun, Resident evil: Understanding residential IP proxy as a dark service, in: IEEE S&P'19, San Fansisco, CA, USA, 2019.

[15] EasyList, EasyList, `https://easylist.to/`.

[16] Shivam Agarwal, BlockListParser, Online article (Jun 02, 2016). `https://github.com/shivamagarwal-iitb/BlockListParser`.

[17] M. Trevisan, S. Traverso, H. Metwalley, M. Mellia, Uncovering the flop of the EU cookie law, arXiv preprint arXiv:1705.08884, 2017.

[18] The EU Internet Handbook, Cookies, Online article (Dec 10, 2018). `http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm`.

[19] S. Aryan, H. Aryan, A. J. Halderman, Internet censorship in Iran: A first look, in: USENIX Workshop on Free and Open Communications on the Internet (FOCI'13), Washington DC, USA, 2013.

[20] P. P. Swire, K. Ahmad, Foundations of information privacy and data protection: A survey of global concepts, laws and practices, International Association of Privacy Professionals, 2012.

[21] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, C. Diaz, The web never forgets: Persistent tracking mechanisms in the wild, in: CCS'14, Scottsdale, Arizona, USA, 2014.

[22] A. Panchenko, F. Lanze, J. Pennekamp, T. Engel, A. Zinnen, M. Henze, K. Wehrle, Website fingerprinting at Internet scale, in: NDSS'16, San Diego, CA, USA, 2016.

[23] G. Tyson, S. Huang, F. Cuadrado, I. Castro, V. Perta, A. Sathiaseelan, S. Uhlig, Exploring HTTP header manipulation in-the-wild, in: WWW'17, Perth, Australia, 2017.

[24] Hola, Hola VPN, `http://hola.org/`.

[25] G. Merzdovnik, M. Huber, D. Buhov, N. Nikiforakis, S. Neuner, M. Schmiedecker, E. Weippl, Block me if you can: A large-scale study of tracker-blocking tools, in: IEEE EuroS&P'17, Paris, France, 2017.

[26] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, T. Holz, We value your privacy... Now take some cookies: Measuring the GDPR's impact on web privacy, in: NDSS'19, San Diego, CA, USA, 2019.

[27] S. Schelter, J. Kunegis, Tracking the trackers: A large-scale analysis of embedded web trackers, in: ICWSM'16, Cologne, Germany, 2016.

[28] Luminati, Monetization SDK, `https://luminati.io/sdk`.

[29] MDN web docs, *X-Forwarded-For*, Online article (July 20, 2018). `https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/X-Forwarded-For`.

[30] S. Huang, F. Cuadrado, S. Uhlig, Middleboxes in the Internet: a HTTP perspective, in: TMA'16, Paris, France, 2017.

[31] Luminati, *X-Forwarded-For* # issue 70, Online article (Nov 17, 2017). `https://github.com/luminati-io/luminati-proxy/issues/70`.

[32] Freedom House, Freedom of the press 2017, Online article (April 2017). https://freedomhouse.org/sites/default/files/FOTP_2017_booklet_FINAL_April28.pdf.

[33] Joshua Roesslein, Tweepy, Online article (Nov 30, 2018). https://github.com/tweepy/tweepy.

[34] A. Gervais, A. Filios, V. Lenders, S. Capkun, Quantifying web adblocker privacy, in: European Symposium on Research in Computer Security'17, Oslo, Norway, 2017.

[35] M. A. Bashir, S. Arshad, W. K. Robertson, C. Wilson, Tracing information flows between ad exchanges using retargeted ads, in: USENIX Security Symposium'16, Austin, TX, USA, 2016.

[36] Google Ad Manager, How exchange bidding works, Online article (2019). https://support.google.com/admanager/answer/7128958?hl=en.

[37] School of Psychology University of New England, Z-scores, Online article (2019). https://webstat.une.edu.au/unit_materials/c4_descriptive_statistics/z_scores.htm.

[38] Citizen Lab, URL testing lists intended for discovering website censorship, Online article (2019). https://github.com/citizenlab/test-lists/.

[39] A. Lerner, A. K. Simpson, T. Kohno, F. Roesner, Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016, in: USENIX Security Symposium'16, Austin, TX, USA, 2016.

[40] Bell Canada, Online advertising program, Online article (2019). https://www.bell.ca/online-marketing.

[41] Canadian Radio-television and Telecommunications Commission (CRTC), The CRTC collaborates with international partners to fight illegitimate online marketing activities, News article (Mar 21, 2018). https://www.canada.ca/en/radio-television-telecommunications/news/2018/03/the-crtc-collaborates-with-international-partners-to-fight-illegitimate-online-marketing-activities.html.

[42] EasyList, Other supplementary filter lists and easylist variants, https://easylist.to/pages/other-supplementary-filter-lists-and-easylist-variants.html.

[43] Google Ads, Understanding Google Ads and AdWords express country restrictions, Online article 2019. https://support.google.com/google-ads/answer/6163740?hl=en.

[44] Michael Carter, pywhois, Online article (Oct 18, 2010). https://pypi.python.org/pypi/pywhois/0.1.

[45] M. Jiang, The business and politics of search engines: A comparative study of Baidu and Google's search results of Internet events in China, New Media & Society 16 (2) (2014) 212–233.

[46] MedCalc, Values of the Chi-squared distribution, Online article (2019). https://www.medcalc.org/manual/chi-square-table.php.

[47] Analytics Help, Understanding PII in Google's contracts and policies, Online article (2019). https://support.google.com/analytics/answer/7686480?hl=en.

[48] Official Journal of the European Union, EU GDPR, Online article (Apr 26, 2016). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN.

[49] ScoreCard Research, ScoreCard Research - Privacy policy, Online article (Dec 19, 2017). http://www.scorecardresearch.com/privacy.aspx?newlanguage=1.

[50] Akamai, Internet connection speeds and adoption rates by geography, Online article (Jun 1, 2017). https://www.akamai.com/us/en/about/our-thinking/state-of-the-internet-report/state-of-the-internet-connectivity-visualization.jsp.

[51] P. Laperdrix, W. Rudametkin, B. Baudry, Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints, in: IEEE Symposium on Security and Privacy'16, San Jose, CA, USA, 2016.

[52] DLA Piper, Data protection laws of the world, Online article (2019). https://www.dlapiperdataprotection.com/.

[53] Freedom House, Freedom on the Net 2017 - Egypt, (Apr. 2017). https://freedomhouse.org/report/freedom-net/2017/egypt.

[54] B. Krishnamurthy, K. Naryshkin, C. Wills, Privacy leakage vs. protection measures: the growing disconnect, in: WTSP'11, 2011.

[55] S. Zimmeck, J. S. Li, H. Kim, S. M. Bellovin, T. Jebara, A privacy analysis of cross-device tracking, in: USENIX Security Symposium'17, Vancouver, British Colombia, Canada, 2017.

[56] F. Wollsén, OpenWPM WebExtension experiment / API, Online article (July 30, 2018). https://github.com/mozilla/OpenWPM-WebExtension-Experiment.

Nayanamana Samarasinghe:

Nayanamana Samarasinghe received Master of Computer Science (MCS) and Master of Engineering (MEng) in Information Systems Security degrees from University of Ottawa, Canada (2006) and Concordia University, Canada (2017) respectively. He is currently pursuing Ph.D. degree at Concordia University, Canada. His research interests include system security, TLS vulnerabilities, web privacy, machine learning and natural language processing.

Mohammad Mannan:

Mohammad Mannan is an Associate Professor at the Concordia Institute for Information Systems Engineering, Concordia University, Montreal. His research interests lie in the area of Internet and systems security, with a focus on solving high-impact security and privacy problems of today's Internet. He is involved in several well-known conferences (e.g., program committee: USENIX Security 2018, ACM CCS 2016, ACSAC 2014, USENIX Security 2010; program co-chair: ACM SPSM 2016; General co-chair: ACM CCS 2018), and journals (e.g., ACM TISSEC, IEEE TDSC, IEEE TIFS).

Conflict of Interest and Authorship Conformation Form

Please check the following as appropriate:

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version. ✓
- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue. ✓
- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript ✓
- The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript: ✓

Author's name	Affiliation
Nayanamana Samarasinghe	Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, Canada
Mohammad Mannan	Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, Canada