Where in the World Are My Trackers?

Mapping Web Tracking Flow Across Diverse Geographic Regions

Sachin Kumar Singh
University of Utah
Salt Lake City, USA
sachinkumar.singh@utah.edu

Robert Ricci
University of Utah
Salt Lake City, USA
ricci@cs.utah.edu

Alexander Gamero-Garrido
University of California, Davis
Davis, USA
agamerog@ucdavis.edu

Abstract

Web trackers are pervasive on the Internet, collecting user data as it flows from the end user's devices to servers. Understanding the physical location of these servers and trackers data flow is essential for assessing privacy risks and ensuring control over user data. While much of the existing research focuses on the Global North, especially the EU where regulations such as the General Data Protection Regulation (GDPR) are in place, the Global South, where most Internet users reside, remains under-studied. We address this gap by collecting measurements from the web browser and the data-plane (IP of responding server & latency) on the same device. A challenge to apply this approach in the Global South is the lower density of observation points in measurement infrastructure. To address this, we build a software suite that automatically collects measurements in various countries from a volunteer's computer. Our suite is interoperable with all major OSes and requires limited user intervention. We apply our method across 23 geographically diverse countries, offering in-depth insights into tracker data flow.

CCS Concepts

• **Networks** → *Network monitoring*; • **Security and privacy** → **Privacy protections**; • **General and reference** → **Empirical studies**; • **Information systems** → **World Wide Web**.

Keywords

Web tracking, Cross-border data flows, Data localization, Network measurement, Under-measured regions

ACM Reference Format:
Sachin Kumar Singh, Robert Ricci, and Alexander Gamero-Garrido. 2025. Where in the World Are My Trackers?: Mapping Web Tracking Flow Across Diverse Geographic Regions. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25), October 28–31, 2025, Madison, WI, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3730567.3764427

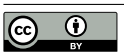
1 Introduction

Trackers are widely present on the Internet [14, 45, 46, 74, 75, 97, 100], recording users' actions across websites. These trackers serve commercial purposes, such as delivering personalized/optimized ads and providing website analytics, but they also pose privacy concerns due to the data they collect [12, 13, 45]. Governments worldwide have attempted to mitigate these concerns using various policy mechanisms, often by placing constraints on the physical locations of web trackers. These constraints, known as *data localization*, limit the transmission of data from the end user's device to tracking servers in other countries. While privacy risks are inherent in both local (or domestic) and non-local (or foreign) trackers, as both types collect user interactions and browsing behaviors, in this study we focus on revealing *non-local* trackers; understanding non-local tracker data flows is crucial as it involves the transmission of data across borders, often to countries with different legal frameworks and data protection standards.

One major challenge is the limited visibility into where trackers are actually hosted *i.e.*, tracker server location. Most existing methods for monitoring trackers and their data flow rely heavily on infrastructure located in the Global North, particularly in Europe [18, 79, 86, 90, 91]. This reliance, coupled with regulations that vary from GDPR to differing degrees depending on the region, complicates efforts to study tracker data flow. Even though a large share of the world's Internet users resides in the Global South, its tracker data flow practices have been researched by limited studies.

Tracker flows are thus not uniformly monitored, leading to significant knowledge disparities across different regions. Addressing this gap requires obtaining accurate browser and IP measurements directly from the countries of interest. These measurements are crucial for determining the tracker server's geolocation, which is fundamental for analyzing the tracker data flow from various countries. Collecting measurements from within the country is essential, as geolocation-based DNS (GeoDNS) [53, 58] and content delivery networks (CDNs) [11, 26] often operate in a location-dependent manner that impacts both the responding server's location and the page content. Consequently, measurements gathered from external networks or other countries may not accurately represent in-country behavior.

Deploying measurement infrastructure in countries of interest is thus essential to study trackers and their localization practices in parts of the world where existing measurement (infrastructure for browser- and IP-based collection) is less dense, especially in the Global South and particularly in Africa. To further this goal, we adopt a set of recently-proposed methods [48] that collect browser and IP-based data from EU users on two measurement platforms. Since the platforms involved have lower density in our regions of interest, we use an integrated approach that combines browser and IP data collection. Our approach relies on volunteers from various countries who use their own machines and Internet connections to perform measurements using our lightweight software suite, which we call ***Gamma***. Our approach addresses geographic and measurement bias introduced by synthetic configurations such as VPNs or Proxies, because these synthetic setups distort latency [94,



This work is licensed under a Creative Commons Attribution 4.0 International License.
IMC '25, Madison, WI, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1860-1/2025/10
https://doi.org/10.1145/3730567.3764427

103] and limit geographic diversity. By using real user vantage points and a validated multi-constraint geolocation framework [48], our measurements more accurately capture where tracker data flows.

We provided the volunteers with *Gamma*, which is easy to use and capable of performing measurements with minimum interaction. Our integrated approach enabled us to collect direct, unbiased data from end-user locations, generating an accurate representation of regional trackers. Additionally, understanding tracker data flow requires precise geolocation of tracker servers' IP addresses, so *Gamma* collected *round trip time* (RTT) server latency, traceroutes for server IPs, and reverse DNS. Combined, this data allows us to determine precise server geolocation, enhancing the accuracy and validity of our findings.

We found that foreign trackers are very common across the world: websites in 91% of the examined countries (21/23) embed trackers hosted in foreign nations. However, there is substantial variation in their prevalence across countries. For instance, while India primarily relies on local servers and trackers, New Zealand depends largely on foreign ones. The destination countries for these trackers generally fall into two categories: neighboring nations with advanced infrastructure and European countries, particularly Germany, France, and the United Kingdom. U.S. organizations dominate the list of observed foreign trackers, and our findings show that they engage in distinct practices depending on the country. Our findings also reveal that factors such as physical proximity and regional dynamics influence inter-country tracking flows. Moreover, countries where these tracking servers are located and their corporate ownership vary widely.

With this work, we make the following contributions:

- We design *Gamma*, a lightweight, highly configurable tool for browser and IP-level measurements (§3)
- We collect data from volunteers' computers across 23 geographically diverse countries spanning every inhabited continent, overcoming infrastructure limitations (§4).
- We analyze tracker data flow practices across 23 diverse countries, examining factors that influence them. (§6).
- We analyze data regulations and policies to assess their impact on tracker data flow (§7).
- Our tool, *Gamma*, along with other artifacts from the paper, is publicly available [2].

2 Background and Related Work

Trackers are ubiquitous on the modern Internet, recording various types of information about users [12–14, 45, 46, 74, 75, 97, 100]. Previous research [48, 62] has demonstrated that these trackers allow data to flow from the end user to non-local tracking servers, located outside the user's country. Governments worldwide are increasingly advocating for mandates requiring user data be stored within a specific nation or region. Regulations have already been implemented in many countries, motivated by concerns over security, privacy, and surveillance [7]. These regulations impose various constraints, such as requiring user consent and establishing rules for transmitting data across national borders. The General Data Protection Regulation (GDPR) [73] is one such regulation applied in the European Union, which restricts the transfer of personal data to non-EU countries that do not meet specific requirements. (Note that we do not study the EU/GDPR in this paper, and only mention it here for context.) However, Internet protocols were not set to track geographic boundaries in their operation, making both compliance with these regulations and audits challenging.

2.1 Limitations of Prior Work and Coverage

While there are studies [48, 62, 110, 124] examining trackers and their flow under GDPR, there is limited research in other regions, despite these other regions having the majority of Internet users. This gap persists due to several factors, including the absence of GDPR-like regulations and a lack of infrastructure necessary for conducting such study. Measurement infrastructures are predominantly located in the Global North, mainly within the EU [18, 79, 86, 90, 91]. As more countries adopt similar regulations (India [56, 57], Malaysia [36]), characterizing tracker flows in countries around the world is critical.

2.2 Synthetic Measurement Points

Studying tracker flows require measurements from an in-country vantage point. Synthetic measurement points-such as cloud machines, VPNs, and Residential IP Proxy-as-a-Service are commonly used alternatives [71, 72, 88, 98]. Overall, these synthetic measurement points suffer from limited geographic coverage, pose a risk of introducing bias, may be flagged by other organizations, and can lead to unreliable or distorted measurements. In this subsection, we describe the limitations of these approaches for our research.

Synthetic approaches have significant limitations in capturing real user experiences, as content providers often treat their traffic differently [25, 98]. Furthermore, these setups are not universally available [69, 112, 113], their performance can be skewed by network architecture and peering relationships, and they can also limit the types of measurements that can be performed.

Synthetic measurement points also introduce geographic and measurement biases. Cloud infrastructure is heavily concentrated in a few regions, often with only a single location per continent, particularly in underrepresented areas like Africa. VPNs present additional issues: they may not be located in their advertised countries [121], suffer from reliability problems, lack transparency [69], and are frequently blocked by countries and websites [27, 29, 30, 32, 47, 81, 82, 85, 116, 120]. Moreover, VPNs introduce variable latency [94, 103], undermining the accuracy of latency-based geolocation techniques [48], which are essential to our study.

Residential IP Proxy-as-a-Service is a "gray" business, which may have illicit activities and compromised IoT devices to serve as proxies [80]. Additionally, these proxies are often found in blocklists and may be labeled as sources of spam [25]. Prior research [98] has observed limitations with residential proxies intentionally blocking some trackers. IP proxy services (e.g., BrightData [15]) provide slightly more geographic diversity, but impose technical constraints that prevent various measurements. For example, many proxies block traceroute and ping, both of which are required in our geolocation method. While past work in Europe was able to combine IP proxies for browser traffic with RIPE Atlas for traceroutes [48], the sparse coverage of RIPE Atlas in the Global South makes this infeasible for our study.

2.3 Global Coverage and Accuracy

In contrast to the limitations of synthetic measurement points, our approach offers both geographic breadth and fidelity. A core strength of *Gamma* is its ability to enable in-country, user-based measurements in regions that are largely invisible to prior work. *Gamma*, by running directly on volunteers' machines over local Internet connections, ensures accurate regional representation and preserves the integrity of latency-based measurements. This allows for low-distortion latency measurements, enabling the use of a validated multi-constraint geolocation framework [48], which has demonstrated 100% precision in identifying *foreign servers*. As a result, *Gamma* is uniquely suited for accurate, large-scale, and representative analysis of tracking flows, particularly in underrepresented regions-providing a more comprehensive view of global tracking practices than previous studies.

3 Design and Usage of *Gamma*

There are two challenges to develop a portable tool for conducting browser and IP-level measurements: users deploying the tool may be using various Operating Systems (OS) and might not have high-end machines. Consequently, we designed *Gamma*, a lightweight tool compatible with a range of operating systems. Despite its minimal resource requirements, it is capable of performing effective browser and IP-based measurements. *Gamma* offers diverse measurement capabilities, allowing users to conduct a wide array of tests at the browser and IP-level. The components of *Gamma* can be divided into three main parts:

(C1) Browser-Level Interaction: Our tool initiates full-fledged browser sessions using the Selenium Webdriver to load specific websites, i.e., Target Websites. *Gamma* supports running measurements across major browsers, including Chrome, Firefox, and privacy-focused Brave. It is capable of saving full webpages, scraping page content, recording HAR files and all network requests during page loads, and downloading associated resources (e.g., CSS, JavaScript, images). Users can customize the number of simultaneous instances they wish to run based on available computational resources and set delays to ensure each page is fully rendered before proceeding.

(C2) Network Information Gathering: This component enriches the data collected during browser-level interaction. It retrieves DNS and reverse DNS information for all captured domains, whether obtained through network requests or hardcoded on the webpage. It queries APIs to annotate domains/hosts with ASN, geolocation, and network/ownership metadata (e.g., IPinfo [63], ipwhois.io [1], RIPE IPmap [64]).

(C3) Measurement Probes: This component of the tool is capable of launching active measurement probes. For instance, it can initiate traceroute probes to domains or IPs identified during browser-level interactions or through the network information gathering process. Furthermore, it supports the deployment of other probes, *e.g.,* ping and TLS using Nmap [87] and Testssl [122], to evaluate network latency, reachability, and security parameters.

For network information gathering (C2) and measurement probes (C3), *Gamma* supports various libraries, such as Scapy [101, 102]. These libraries are integral for capturing and analyzing network traffic and conducting network probes. However, we encountered challenges with some libraries not being universally compatible across different operating systems. For instance, majority of features of Scapy don't work on Windows OS. To overcome this, we added functionality to *Gamma* that uses OS-specific commands and tools to perform various measurements. For example, it utilizes the `traceroute` command in Linux and the `tracert` command in Windows. However, this approach introduced output variability: these commands produce output in different structures. To address this, we developed additional functionality that normalizes the output into a consistent format, regardless of the method used. For instance, *Gamma* produces an identical structure JSON file with hop and RTT information for `traceroute` and `tracert`. This ensures a standardized output format and reduces variability.

In addition to the functional components (C1, C2, C3), *Gamma* is designed with a strong emphasis on *portability*, achieved by incorporating OS-specific functionalities and by supporting a variety of browsers, allowing users to conduct comprehensive studies and comparisons across different browsers without being restricted to a specific platform.

For this study, *Gamma* utilizes only the following components: an isolated Chrome browser instance (C1), DNS and reverse DNS of network requests (C2), and traceroutes to all resolved IPs (C3).

3.1 Tuning *Gamma*

For this study, we configured *Gamma* to launch isolated Chrome browser instances to load websites and perform three main functions, each building on the previous step: record the domains to which the browser (instance) sends network requests, resolve forward and reverse DNS for all domains present in these network requests, and initiate a traceroute to all resolved IP addresses. Given that our volunteers may not always have access to high-end machines, we generally configured our tool to run in a single-thread mode. Additionally, we used Chrome browser with a wait time of 20 seconds to completely render the website, which is typically double the time required for most websites to render fully [17]. Since we are aiming to access ≈100 websites on volunteer machines, therefore, using a large wait time is not feasible. We have also set a hard time constraint of 180 seconds, which serves to address any incidents where the browser instance may become non-responsive for any reason. In such cases, our tool will automatically terminate that browser instance and proceed to the next website.

3.2 Picking Target Websites

Target websites (T_{web}) are the set loaded by *Gamma* to perform measurements. Our target websites (T_{web}) comprise two categories: top regional websites (T_{reg}) and regional official government websites (T_{gov}).

We used region-specific websites for two reasons: first, our study focuses on specific countries (see §5 for a list of countries and our rationale for the sample); second, global rankings often overlook certain types of websites and regions [95]. We selected 50 top regional websites from similarweb [106] for each country, a strategy used in previous research [16, 96, 130]. For some countries, similarweb does not have regional website rankings. We evaluate two other potential sources, semrush [104] and ahrefs [4], by analyzing the overlap in the top 50 websites for 58 different countries across lists available from similarweb, semrush, and ahrefs. We select these

58 countries because all three sources provide complete lists of sites. We then measure the overlap percentage between each alternative source and similarweb. Semrush shows a 65% overlap, indicating a closer alignment to similarweb's lists compared to ahrefs, which only showed 48% overlap. Thus, we used semrush's rankings for countries where similarweb was unavailable.

The 50 top regional websites include a diverse range of sites, including news outlets, e-commerce platforms, and local service providers. We removed all adult sites and websites banned in each country. For each country, we got a unique T_{web}; only two websites, *google.com* and *wikipedia.org* were common across all countries. Additionally, seven more websites (*instagram.com*, *youtube.com*, *facebook.com*, *openai.com*, *twitter.com*, *whatsapp.com*, and *linkedin.com*) appeared among the top regional websites in at least two-thirds of the countries for which we recorded data.

Official government websites are crucial for providing public services, sharing information, and may contain sensitive information. Government sites are often used by the regional population and are highly trusted by the public [128]. Past studies have found commercial tracking tools, trackers and cookies on government websites and apps [54, 97, 109, 118, 130]. Because many public services are only available through official portals and government sites, citizens often have little choice but to use these sites. Studying these government sites therefore reveals the tracking exposure of real users and the in-region practices. To compile a list of government websites, we used the Tranco list [92, 117] (January-2023) and filtered out websites by utilizing government-specific Top-Level Domains (TLDs). These TLDs are only registered and utilized by national governments, *e.g., .gov.au* is used by official government websites in Australia. Some countries have more than one TLD; for example, Argentina uses both *gob.ar* and *gov.ar*, we considered multiple TLDs while compiling the list, similar to previous studies [54, 66]. We selected 50 government websites from the Tranco list for each country. For countries with fewer than 50 government website in Tranco we scraped Google search results for government TLDs and added them to T_{gov}.

We incorporate both government websites and regional websites in our analysis because they collectively offer a comprehensive view of how the Internet serves the local population. Volunteers are provided with the T_{web} list and can opt out from accessing any number of the websites.

3.3 Recruitment

We used various methods and channels for recruiting volunteers for the data collection. We reach out to individuals within our existing personal networks. We also recruited volunteers by posting a recruitment message on social media platforms. The message includes details of the study, a call for volunteers from various countries who are willing to participate and execute *Gamma* for data collection. Interested individuals are encouraged to reach out to us via the provided email. Additionally, we employed snowball sampling to recruit further volunteers. Identified interested participants were further provided with more details via email. This included detailed information about the study, a consent document, and tuned *Gamma* along with instructions for its execution. The email requested volunteers to review the consent document, which offered in-depth details about our study. It explained what data we are recording, how the data will be stored, their role in the study, and emphasized that their participation is entirely voluntary and that they could withdraw at any time. Volunteers were also informed that they could opt out of specific components of the data collection if they wished *e.g.*, they could opt out of accessing any website from T_{web}.

We encouraged participants to ask questions about the consent document or the tool and offered to schedule meetings for clarification or assistance with setting up the tool. After going through the consent document, if the volunteers decided to proceed, they follow the instructions to set up *Gamma* and begin the data collection, which was expected to take few hours. We advised volunteers to complete the experiment in a single session to minimize variability. However, volunteers can also run it in chunks, as *Gamma* is designed to resume from where it was last stopped. With 22 volunteers we were able to perform our study in 23 diverse countries (one volunteer recorded data for two countries).

3.4 Country Selection

This study relies on volunteer participation to gather data from diverse global regions. The choice of countries is shaped directly by the location of the volunteers who responded to our outreach. Our aim is to gather data from a globally-diverse group of countries. However, the diverse representation achieved through these volunteers provided rich insights into tracker behavior, with 4 countries from Africa, 11 from Asia, 2 from Europe, 2 from North America, 2 from Oceania, and 1 from South America.

Our volunteer-driven approach brought participants from Global North (developed) nations like the United Kingdom, Canada, and Japan while also enabling data collection in Global South (developing) regions like Rwanda, Uganda, and Azerbaijan. The latter countries, which are less frequently studied in Internet tracking and privacy research, provided an invaluable perspective on how trackers operate in regions with varying levels of infrastructure and regulatory frameworks. This diverse representation allows our findings to reflect not only well-known trends in the Global North but also the unique dynamics seen in countries across Africa, the Middle East, and South Asia.

3.5 Ethics

Volunteer Privacy: We only utilize volunteer machines as vantage points and we do not access any pre-exsiting data on these machines. All the browser instances used are also isolated and do not access volunteers' browser account nor history.

Data collection and handling: We follow the best practices for recording and analyzing data, including data minimization [60]. We strictly record data that is necesarry for our study. Our method does not record any identifiable information about the participant, except for the IP address. However, as most network providers utilize dynamic IP assignment and NAT [42], it is not directly straightforward to link the IP to the volunteer. As an additional precaution, after completing our analysis (section 4), all volunteers IP addresses are anonymized within the dataset to ensure confidentiality.

Volunteer accommodations: Our volunteers are based in various countries, each with potentially different preferences and we

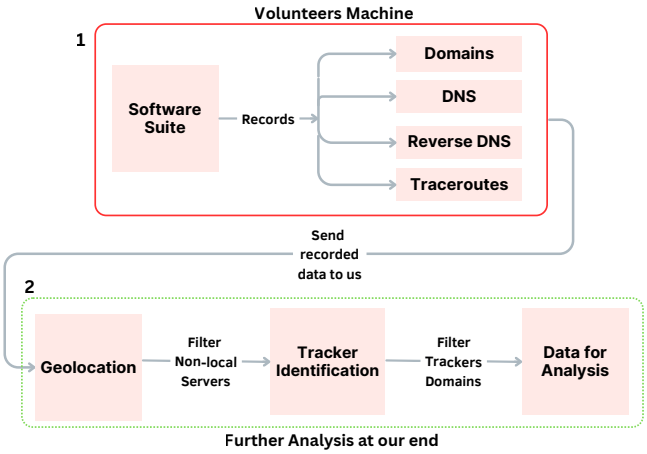


Figure 1: Process diagram of our method. Box 1 (top) shows the volunteer machine running our software suite. Box 2 (bottom) shows the further analysis pipeline.

make every effort to accommodate them. For instance, one volunteer opted out of launching traceroutes, and we ensured compliance. Another volunteer requested a demonstration by running the code for a specific website to better understand the experiment before proceeding with the full data collection, which we provided.

Ethics Statement: Based on our methodology and software suite, this work does not raise any ethical issues, and it has been acknowledged by the university's Institutional Review Board (IRB) as Non-Human Subject Research (NHSR).

4 Measurement Methodology

Figure 1 shows the various steps and the flow of our method. Box 1, on the top side of the flowchart, represents the volunteer machine, and Box 2, on the bottom, displays the various steps that we subsequently perform after getting the recorded data.

Volunteers Machine (Box 1): Volunteers are provided with the software suite and detailed instructions for installing dependencies and executing the suite for performing measurements. For each website from Target Websites set (T_{web}), the tool records the domains, DNS, reverse DNS and the traceroute to IPs. Once the tool execution is completed, the volunteer sends us all this recorded information. We ask the volunteer to disclose their city and the tool also logs the volunteer's IP. We use the volunteer city location and network they connect to (inferred from their IP) to perform precise geolocation of tracker servers (§4.1).

Further Analysis (Box 2): Once we receive the data from the volunteer, we infer the location of responding servers (Section 4.1) and tracker identification (Section 4.2) for further downstream analysis.

4.1 Precise Geolocation of Domains

Geolocating servers is essential for assessing tracker flow, but it is not an exact science. Various commercial and non-commercial databases (e.g. MaxMind [78], NetAcuity [44], DB-IP [37] IPinfo [63], RIPE IPmap [64]) have been used by researchers for IP geolocation. However, studies have shown they are not fully reliable [31, 50, 59, 76, 93, 105]. We use a recently-proposed method [48] that combines geolocation databases, source-based constraints, destination-based constraints, and DNS records to infer where servers are located.

Previous research has identified RIPE IPmap as the most reliable service for IP geolocation [38, 62] and RIPE IPmap has been extensively used in previous studies [20–22, 38, 49, 62, 76]. We used RIPE IPmap to geolocate all the IPs obtained from DNS lookups of network requests on volunteer machines. We categorized the data by geolocation; IPs located outside the volunteer's country were classified as *Non-local*, while those within the country were classified as *Local*. We conduct additional tests to confirm that servers are *Non-local* and reduce the impact of IPmap's potential inaccuracies on our results. First, *source based constraint* that involves analyzing response times from a volunteer machine to the tracker server. Secondly, the *destination based constraint*, which assesses the response times from a RIPE Atlas probes to the tracker server. Lastly, the *reverse DNS constraint* that utilizes reverse DNS data based on hostnames [77], where available, to rule out IPs with potentially incorrect geolocations.

Speed of Light Physical Constraint in Cable: In analyzing network latency, it is crucial to consider physical constraints. Specifically, the observed speed of data transmission in our measurements based on round trip time from traceroutes should not exceed $\frac{2c}{3}$(c represents the speed of light), i.e., 133km/ms [67], based on transmission rates in fiberoptic cable. We use this Speed Of Light (SOL) physical constrain in our source and destination based constraints.

4.1.1 Source based constraint. After identifiying non-local servers (IPs) using the geolocation form RIPE IPmap, we analyzed the latency (round trip time) recorded through the traceroute launched by our tool on volunteer's machine. To obtain accurate latency measurements, we subtracted the recorded last hop time from the first hop, only if first hop time is available and is smaller than last hop, if not then we consider the last hop as latency. We also do this for our destination-based measurements. This subtraction removes the delays caused by the local network, making our latency measurements more accurate. We discarded all non-local servers for which the traceroute did not reach the destination. We also discarded all the measurements that fail the SOL constraints.

We then compare this data to statistics of latency previously observed between the geographical location of the volunteer and the server. We obtained these statistics from Verizon [119]. For cases where Verizon does not offer data, we used latency statistics from WonderNetwork [126]. As a conservative measure, we discard all non-local servers/IPs where the latency in our observations is less than 80% of the latency statistics.

Traceroutes in Egypt, Australia, India, Qatar and Jordan. Our volunteer in Egypt choose not to launch traceroute probes. In the other four countries the traceroute probes failed. We do not know the exact cause, though local network configuration or firewalls are potential reasons. In these countries we find RIPE Atlas probes, and launch traceroutes to all server IPs from the volunteers' data, within a few days following the initial data collection. We select probes as close as possible to the volunteer's city and on the same network, where feasible. In two cases the RIPE probe is in a nearby country: Saudi Arabia for Qatar and Israel for Jordan.

4.1.2 Destination-based constraint. To verify the geolocation of a server, we used *destination traceroutes*. These are traceroutes from a RIPE Atlas probe that is located in the same country as the non-local server according to the RIPE IPmap geolocation data. We choose the probe in the same city when available. We discarded all non-local servers for which the destination traceroute did not reach the destination, and all measurements where the SOL constraint is violated.

4.1.3 Reverse DNS based constraint. We examine the reverse DNS records for each server IP. Although not always available, reverse DNS can offer valuable insights about the IP's geographical location [48, 77]. We utilized these clues to identify and eliminate non-local servers that do not match the RIPE IPmap geolocation. Specifically, we conducted a manual inspection of all reverse DNS entries for non-local servers. If the reverse DNS suggested a geolocation that contradicted the geolocation information provided by the RIPE IPmap, we excluded it from our data. Conversely, if the reverse DNS did not provide clear geographical hints, the servers are retained in our data. For example, for T_{web} in Pakistan, a few Google-owned IPs were geolocated to Al Fujairah City, United Arab Emirates by RIPE IPmap but the reverse DNS information showed evidence for Amsterdam. Similarly, for T_{web} in Egypt, Google-owned IPs were geolocated to Germany but reverse DNS information suggested Zurich, Switzerland, so we discarded them.

4.2 Tracker Identification

To conduct the initial filtering of domains associated with advertisement and tracking, we used filtering lists easylist [40] and easyprivacy [41]. These lists are commonly used by ad-blockers and in prior research [3, 5, 24, 62, 65, 108, 125, 129]. They are designed to block ad scripts, ad images, analytics scripts, fingerprinting, email tracking, among other tracking and ad related elements [39]. We also utilized regional ad and tracker lists where available [51, 52, 107].

Given our emphasis on non-local trackers, we compare all the non-local domains with these lists to identify ad and tracker domains. While these lists are comprehensive, they may not capture all regional ad and tracking domains. Therefore, for the remaining non-local domains, we conducted a manual inspection using WhoTracksMe [123], which has also been utilized in previous research [16, 33, 99], along with a cursory Internet search to determine whether they are trackers. For example, we manually labeled *the-ozone-project.com*, a digital ads platform not found in the lists. Using lists and manual inspection, we identified 505 (441 from lists, 64 manually) unique non-local ad/tracking based domains including registrable domains (eTLD+1) such as *googletagmanager.com*, *doubleclick.net*, and *googleapis.com*, plus a few full hostnames (FQDNs) like *693...safeframe.googlesyndication.com*.

5 Data Collection

We collected data in 23 geographically diverse countries, in both the Global North and the Global South. All these countries are represented on the x-axis of Figure 2. Notably, the majority of these countries have not been studied before, particularly in the context of tracker data flow.

The Target Website list (T_{web}) provided to the volunteers consisted of 2005 websites, in total. The volunteer opt outs were minimal: only 0.99% of the websites. Following the opt outs, we had 1987 websites, 1522 of which were unique, across all T_{web}. For the rest of the paper T_{reg}, T_{gov} and T_{web} (T_{reg}+T_{gov}) will refer to the lists of regional, government and target websites respectively, that remained after the volunteer opted out. Figure 2(a) illustrates the number of T_{reg} and T_{gov} websites in T_{web} for each country (§ 3.2). In some countries, such as Lebanon, Russia, and Algeria, only a small number of government websites appeared in our input datasets.

Figure 2(b) illustrates the percentage of T_{web} that were successfully loaded by our tool. Our tool successfully loads and records data for over 86% of the websites across countries. However, our tool only recorded data for 64 and 56 percent of the websites from T_{web} for Japan and Saudi Arabia, respectively. While our tools do not determine the causes of these failures, we speculate that factors that may have contributed to this relatively lower coverage include the quality, speed, and stability of internet connections.

During all our experiments, our tool recorded ≈26K domains (≈5K unique) across all countries by collecting all the network request made by the browser while loading the T_{web}, which resolved to ≈9K unique IP addresses. These requests were sent both to the initial target sites, as well as any additional domains loaded by them as reflected in the network requests.

We launched ≈27K source traceroutes combined from volunteer (≈25K) machines and from RIPE Atlas. For volunteers, the average was ≈1.4K, with the highest number of traceroutes launched in the USA (≈2.2K), followed by Canada (≈2K), the UK (≈1.9K), Thailand (≈1.8K), and Argentina (≈1.7). The lowest numbers were observed in Lebanon (≈1K), Taiwan (≈0.8K), and Saudi Arabia (≈0.4K). We also launched ≈3.4K destination traceroutes in more than 60 different destination countries from RIPE Atlas.

In our data, we identified ≈14K domains that were non-local across all countries (out of ≈26K). After applying our SOL constraints (section 4.1.1, 4.1.2), we got ≈6.1K domains and after applying reverse DNS constraints (section 4.1.3), we were left with ≈4.7K non-local domains for further analysis. Of these, ≈2.7K were associated with trackers.

During our analysis we also noticed that the chrome webdriver used by selenium was generating some google services requests while loading website; this has also been observed previous research [23]. We removed these requests from our data before doing further analysis.

6 Analysis

After filtering the data based on precise geolocation constraints, we analyze the prevalence of non-local trackers, identify central hubs for tracking servers, and ascertain which organizations trackers belong to. We refer to countries where measurements were taken as *"sources"* and those where the servers are geolocated as *"destinations."*

Specifically, we address the following research questions:

- **RQ1:** *How prevalent and how heterogeneous are non-local trackers across countries and site types (government and regional)?* (§6.1, §6.2)

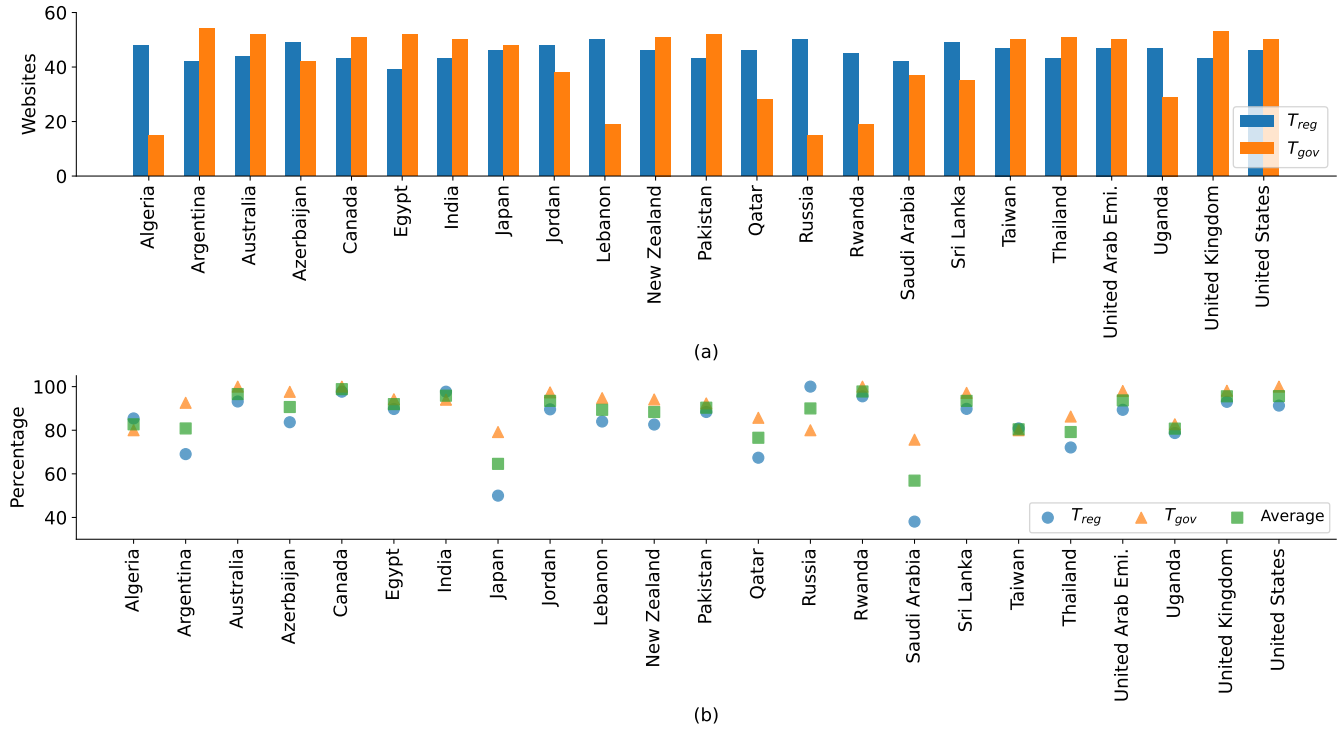


Figure 2: Plot (a) shows the number of regional (T_{reg}) and government (T_{gov}) websites within the Target Website (T_{web}) list. Plot (b) shows the percentage of websites for which our tool successfully loaded and recorded data.

- **RQ2:** *Which countries serve as hubs for hosting non-local tracking infrastructure, and how are tracking flows distributed globally?* (§6.3, §6.4)
- **RQ3:** *Which companies operate the observed non-local trackers, and how geographically diverse is their hosting?* (§6.5, §6.6)
- **RQ4:** *Are first-party non-local trackers a significant component of observed cross-border tracking?* (§6.7)
- **RQ5:** *Do data localization regulations influence the prevalence of non-local trackers?* (§7)

6.1 Prevalence of Non-local Trackers

Our first analysis, shown in Figure 3, looks at the *fraction of websites* that embed *at least one* non-local tracker. As observed in previous studies [14, 45, 46, 74, 75, 97, 100], it is expected to see most websites are embedded with *some* trackers, including government websites. The presence of non-local trackers varies significantly across different countries and across their regional and government websites.

On average, 46.16% of regional websites host non-local trackers, with a standard deviation (σ) of 33.77%, indicating that nearly half of the websites transmit user tracking data outside their country. The high σ shows high variability in the percentage of websites with non-local trackers across different countries. This percentage range extends from 0%, where countries like Canada, India, and the USA have no regional websites with non-local trackers, to countries where almost every regional website embeds non-local trackers. For example, Rwanda shows high percentage, with 93% of Rwanda's regional websites embedding non-local trackers. Other notable examples include Qatar with 83%, Azerbaijan with 82%, and New Zealand with 81%. Similarly, for government websites, 40.21% embed non-local trackers, again with a high standard deviation (σ = 31.5%). Canada, Russia and the USA are the only countries with no government websites embedding non-local trackers, whereas New Zealand (85%) and Uganda (83%) have the highest percentages.

In general, there is a correlation between the presence of non-local trackers on T_{reg} and T_{gov} websites (0.89 Pearson correlation coefficient), indicating that the way countries handle tracking on regional and government websites tends to be similar. There are, however, notable instances with large differences. For example, in Australia, 12% of T_{reg} websites embed non-local trackers compared to 1% of T_{gov}. Similar trends are observed in Azerbaijan (82%, 65%), Qatar (83%, 62%), Russia (16%, 0%) and Rwanda (93%, 31%). It is generally the case that the percentage of regional websites embedding non-local trackers is greater than or equal to that of government websites, but there are exceptions, such as the United Arab Emirates (26% for regional and 40% for government websites), Uganda (67%, 83%) and Taiwan (5%, 10%).

6.2 Non-local Trackers Per Website

Our next analysis looks at *how many* non-local tracking domains are present on each website. In the context of this work, "domain" refers to the segment of a URL that precedes the path. This includes the top-level domain, second-level domain, and any subdomains

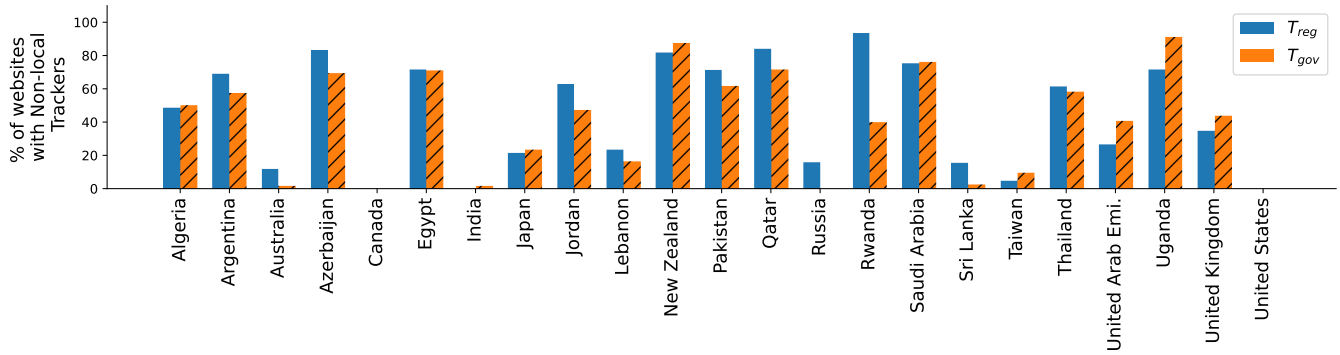


Figure 3: Percentage of Regional (T_{reg}) and Government (T_{gov}) websites with non-local trackers.

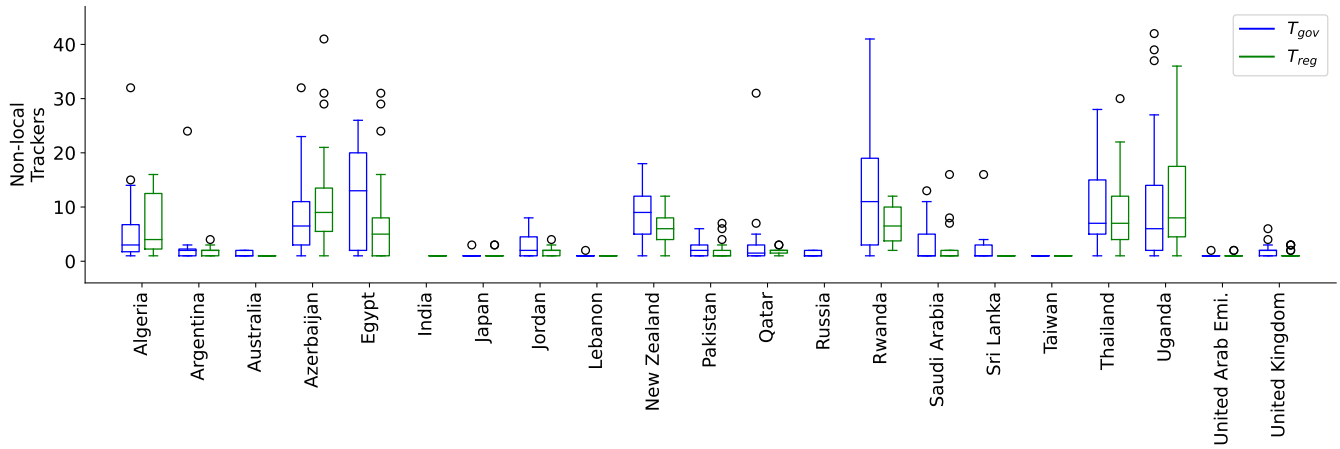


Figure 4: Distribution of non-local tracker domains per-website across Regional websites (T_{reg}) and Government websites (T_{gov}). Each box summarizes the per-website counts within a country.

(*e.g.*, *www.a.b.c.com* and *www.q.w.c.com* are considered different domains).

Figure 4 shows box plots visualizing the distribution of non-local tracker domain counts across various countries. The median number of tracking domains per website is less than ten in most countries. For most countries, the distribution has a positive skew, indicating a concentration of low values, with relatively few high values. Similar trends are observed in T_{reg} and T_{gov} in most countries, illustrating similar patterns of tracker usage in government websites.

Some countries do have notably anomalous trends. Only New Zealand's data exhibits a normal distribution, suggesting a more even spread of tracking domains across websites than typical elsewhere. T_{reg} in countries such as Egypt, Rwanda, Uganda and Thailand exhibited wide interquartile ranges, suggesting a significant variability in the number of trackers across T_{reg}. In countries like Argentina and Qatar, the vast majority of data points are low, with the number of tracking domains predominantly ranging between 1 and 3. These countries also feature outliers, which stand distinctly apart from the rest of the distribution, showing deviations in non-local tracking domain usage among some websites (Appendix A has more details).

Websites in countries such as Jordan, Egypt, and Rwanda possess a high number of unique non-local tracking domains, with averages of 15.7, 12.1, and 13.3 per website, respectively. These countries also show high standard deviation (σ 12, 8.5, 11.39), indicating high variability among the unique tracking domains across various websites (T_{web}). Other countries exhibited low averages, indicating few non-local trackers per website: Australia, Taiwan, Argentina, Lebanon, the UK, and Russia had low averages of tracking domains (1 to 3), with low σ.

Several countries exhibited outliers in the distribution of non-local tracking domains, indicating that while most websites typically host a moderate number of trackers, a few embed an unusually high number of non-local trackers. Upon manually inspecting data from these websites, we found that most of them contain a significant number of tracking domains from well known large tracking organizations, such as Google, Facebook, Twitter, and Amazon. For example, in Azerbaijan, Youtube embeds 32 tracking domains,

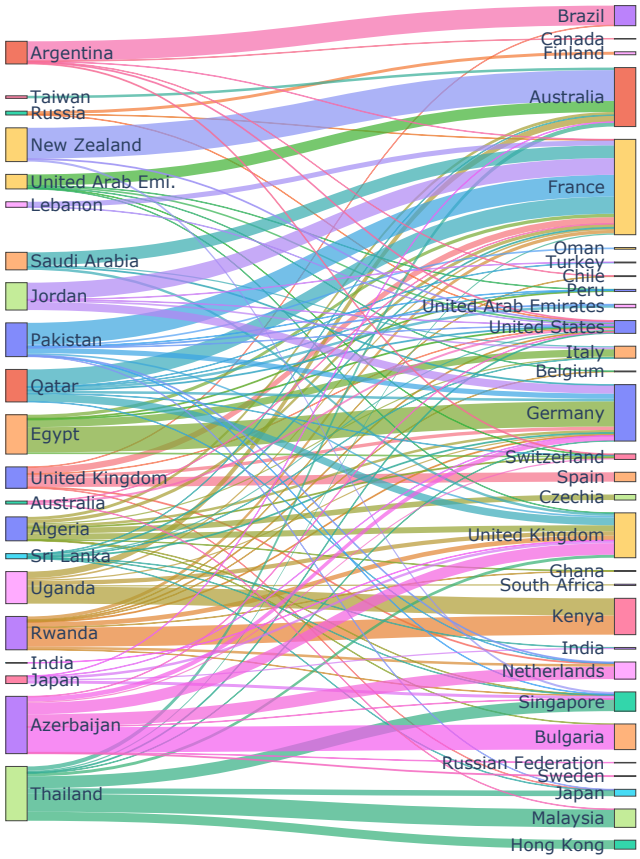


Figure 5: Non-local tracking flows from source countries (left) to destination countries (right).

all associated with Google, comprising a variety of second-level domains (*e.g.*, *googleapis.com*, *doubleclick.net*, *google-analytics.com*, *googleadservices.com*, *googlesyndication.com*). Similar patterns are observed for websites in Azerbaijan (*dost.gov.az*, *edu.gov.az*) and other countries as well.

While most outlier websites primarily host trackers from major tracking networks, there are exceptions where websites feature a diverse array of non-local trackers from various third-party providers apart from major tracking networks. For instance, the regional website *manoramaonline.com* in Qatar not only includes trackers from well-known tracking networks like Google and Amazon Ads but also from Twitter and several other advertising services such as *dotomi.com* and *smaato.net*. Similarly, in Uganda, the website *koora.com* is another notable outlier. Besides trackers from major tech companies, it uses a range of third-party tracking services, including *spot.im*, *scorecardresearch.com*, *33across.com*, *open-x.net*, *360yield.com*, and others.

6.3 Locations Of Tracking Servers

We now investigate the destination countries where non-local trackers are hosted. Figure 5 shows a flow diagram for non-local trackers, with the thickness of each flow representing the number of websites in the source country that transmit data to trackers hosted in the destination country. Among websites with non-local trackers (T_{reg} and T_{gov}), 43% of the websites use *at least one* tracker hosted in France. This is followed by the United Kingdom (24%), Germany (23%) and Australia (23%). Kenya hosts trackers for 14% of analyzed websites, while trackers in other countries are used by less than 10% of websites.

A closer look into the data reveals that some destinations have high flow due to *single* sources. For instance, a significant proportion of New Zealand's tracking is directed towards Australia. If data from New Zealand is excluded, the percentage of websites with tracking flows to Australia decreases from 23% to 11%. Similarly, websites with tracking flows to Malaysia represent 7% of the total, but this number dramatically drops to just 0.16% when contributions from Thailand are excluded. We observed similar trends across several other countries. Bulgaria, Hong Kong, Japan, Brazil, and Finland are destinations for moderately high tracking flows that are sourced from a single country.

We also observed a regional pattern where a destination country has high tracking flows largely because websites from *a few* nearby source countries predominantly use trackers hosted in that destination. This is particularly true for Kenya, which receives significant tracking flows from 14% of the websites across all T_{web}—most of this flow is from T_{web} from neighboring African countries Uganda and Rwanda.

One might expect to observe similar behaviors across different regions, where a few countries act as tracking hubs for nearby websites. However, the dynamics in Asian source countries, such as Pakistan, Sri Lanka, India, and Thailand, exhibit the opposite behavior. We observed that all the major tracking networks have servers in India. Almost all Indian T_{reg} and T_{gov} show no non-local tracker flow, while Sri Lanka has minimal non-local tracker flow activity. On the other side, both Pakistan and Thailand exhibit substantial tracker flows, but with distinct patterns. The majority of tracker flows from Pakistan are directed towards France and Germany, followed by significant flows to the UAE and Oman. On the other hand, Thailand's trackers predominantly flow to Malaysia, Singapore, Hong Kong, and Japan. This different behavior can be because of availability of nearby infrastructural options. For instance, Singapore serves as a major hub for several tech giants and is geographically close to Thailand, making it a prominent destination for trackers. Similarly, the proximity of the UAE and Oman to Pakistan may explain the high volume of tracker flows to these countries.

We have also observed that some destinations receive tracker flows from a significantly broader range of source countries. France and the USA each receive tracking flows from 15 source countries, while Germany and the United Kingdom receive flows from 13 and 12 source countries, respectively. Specifically, our data shows that only 5% of websites direct at least one tracking flow to the USA, which is substantially lower than France (43%), the United Kingdom (24%), and Germany (23%). This disparity is particularly noticeable between T_{reg} and T_{gov} . For T_{reg} websites, France, Germany, and the USA received tracker flows from 13, 12, and 15 countries, respectively. However, for T_{gov} websites, France and Germany received flows from 10 countries each, whereas the USA received flow from only one country, the UAE. This significant variation underscores

the relatively minor role the USA plays in receiving website tracking flow, as compared to France and Germany, in our sample.

6.4 Flow Across Continents

Figure 6 shows the flow across continents. Websites in various countries in Africa possess non-local tracker flow to both Africa and other regions. The majority of these non-local trackers in Africa are hosted in Europe, followed by Africa itself. Asia shows similar trends, with the majority of non-local tracker flow directed to Europe, followed by Asia itself. For countries in Oceania, most of the non-local tracker flow remains within Oceania, largely due to the flow from New Zealand to Australia. In South America, the majority of the tracker flow stays within the continent. However, unlike all other continents that exhibit *inward* tracking flows, Africa stands out as the only continent with no inward tracking flow from any other region.

Only Europe receives significant inward non-local tracker flows from all other continents. This underscores Europe as a *central hub* for tracking data collection. It is well-known that most of Europe restricts outward data flows due to GDPR regulations, and our observations also suggest that countries in North America (USA and Canada) do not have significant non-local outward tracker flows. Taken together, these findings suggest that Europe and North America primarily receive tracking data from around the world and they do not transmit tracking data. Additionally, given the volume and diversity of inbound tracker flows from all continents, Europe clearly emerges as the central hub for global tracking data aggregation.

6.5 Flow To Organizations

We performed manual inspection of all the *organizations owning non-local tracking domains* using whotracksme [123] and Internet search. Figure 8 shows these owners organizations. Not surprising, the majority of domains belong to Google. There are relatively few tracking organizations which are only used by websites in a specific country. For example, we only found trackers owned by Jubnaadserve, onetag, optad360 embedded in websites in Jordan. Similarly, there were some tracking organizations only used by websites in Qatar, the United Kindgom, Rwanda, Uganda and Sri Lanka. We also identified ≈70 companies that own all the non-local tracking domains. Of these, 50% are based in the USA, followed by the UK (10%), the Netherlands (4%) and Israel (4%).

We also performed AS-level lookups on non-local tracker's IP addresses, which showed that a majority of tracking networks are hosted within AWS or Google Cloud. Looking closely at the data from source countries, we identified 50 trackers hosted on Amazon Web Services (AWS) and 5 on Google Cloud. Many trackers encountered by volunteers from Rwanda and Uganda are hosted on IP addresses owned by Amazon in Nairobi, Kenya. These include trackers from various providers like SoundCloud, Spot.im, Snapchat, ScorecardResearch, and Lotame. Interestingly, at the time we took our measurements, AWS did not have a region in Kenya, though one had been announced to open in the future [68, 114]. There is, however, a Cloudfront Edge location in Kenya [10] that could account for this traffic.

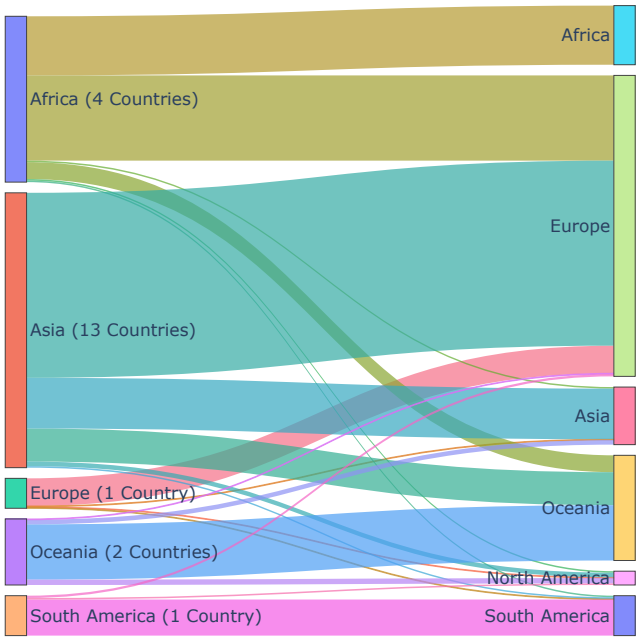


Figure 6: Non-local tracking flows across continents. Left nodes represent the continents of the source countries and right nodes represent the continents of destination countries hosting the tracking domains.

6.6 Domains By Hosting Country

We analyzed the distribution of *specific non-local tracking domains* in relation to the *destination countries* in which they are hosted. Kenya hosts the most with a total of 210 domains, followed by Germany (172) and France (92). In contrast, the USA, which is typically known for its tech infrastructure and has the largest number of datacenters [34, 111], only hosts 16 non-local tracking domains. Other countries such as Belgium, Ghana, Turkey host only one tracking domain.

While the usual Global North countries are present among the most prevalent hosts of tracking domains in our data, our findings show that countries in the Global South also host substantial numbers of trackers. Besides Kenya, Malaysia also hosted 89 tracking domains that are utilized by websites in other countries, showing their roles as tech hubs in Africa and Southeast Asia, respectively. Figure 7 shows the variation in the distribution of hosting countries of non-local tracking domains that are utilized by the T_{web} while loading in various countries.

6.7 First and third party non-local trackers

We performed analysis to examine whether the embedded non-local trackers in T_{web} are first-party or third-party. A tracker is deemed first-party if it belongs to the same organization as the website [19]; otherwise, it is considered third-party. Among 575 websites with non-local trackers across all source countries, only 23 were found to embed first-party non-local trackers. Notably, about 50 percent of these websites are associated with Google, registered under various

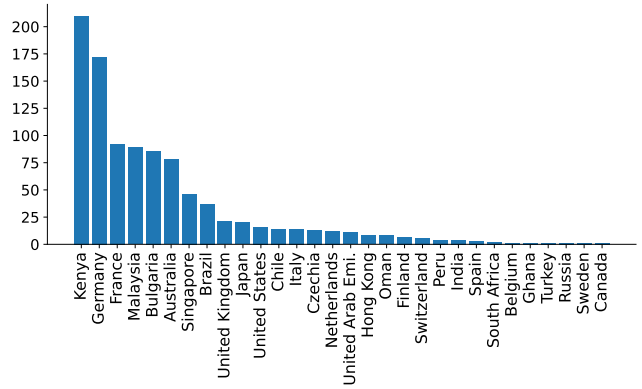


Figure 7: Hosting country distribution of non-local tracking domains by measurement country. Countries on the x-axis are ordered from highest to lowest total count of non-local tracking domains.

country-specific, top-level domains (*e.g.*, *google.com.eg*, *google.co.th*, *google.com.qa*, *google.jo*). The other websites with first-party trackers are affiliated with Facebook, Twitter, Booking.com, BBC, Yahoo, and Microsoft.

7 Discussion

Tracking flow is likely influenced by factors beyond geography: In some cases, the locations of trackers follows geographic proximity, as we saw with Kenya, where a majority of websites from close proximity countries send tracking data there. Kenya seems to have the requisite infrastructure and it is also well connected with submarine cables [70, 84], which makes it a preferred location for hosting trackers in the region.

The influence of geographical proximity on tracking flows is not always straightforward. For example, Egypt shows a significant tracking flow to Germany, and most of these trackers are linked to Google. Despite Google's tracker and ad services servers being hosted in nearer Italy and France, the majority of this traffic from Egypt is routed to farther Germany. We do not know the cause, but speculate that potential reasons are reduced network latency, the presence of infrastructure for specific APIs or German companies with historical ties to Egypt, or reduced downtime.

Sri Lanka, despite close proximity to India and a cable link [83], does not direct significant data traffic towards it. For T_{web} in Sri Lanka, only one tracker from *adstudio.cloud* routes to India. Conversely, the majority of Yahoo trackers in Sri Lanka are directed towards Japan. This may be because Yahoo shut down its news website in India in August 2021, a move driven by new regulations that restrict foreign ownership of media companies in India [89, 115, 127].

Similarly, despite being geographically close to India, Pakistan does not direct tracking traffic towards it, even though both countries have landing points on IMEWE (India-Middle East-Western Europe) [61] fiber cable network. Further, major tracking and advertising providers have servers hosted in India. The lack of traffic flows might be caused by the diplomatic tensions between the two nations.

Global Policy Regulation: We investigate whether public policy impacts the rates of non-local trackers. These laws have proliferated around the world in recent years. Data localization policies are typically enacted by regulating data transfers. For example, in Algeria's Law 18-07 [6], which requires tech companies to obtain government approval before transfering personal data abroad. Argentina, conversely, allows data transfers [9] only to pre-approved countries (e.g., EU members), a similar approach to the EU's GDPR; transfers to other countries are limited to specific types of transactions. Yet in other countries, such as the US, there are stringent protections only for certain types of data (including health records), whereas other transfers are allowed if US data protection laws are followed overseas. Similarly, countries including Australia, Canada and Japan do allow data transfers overseas but still require data to be processed in compliance with domestic regulations.

Table 1 shows our findings. We group data localization regulation in four types, and sort the table based on the regulation's strictness. (The laws in India and Pakistan are not yet in effect, whereas Thailand's was enacted after our data collection was over). Note that we do not know if each company hosting a tracker has obtained the necessary permits in each restrictive country, so the existence of foreign trackers does not necessarily mean that the company is breaking the law. [1] Still, as evidenced by the second and last columns in Table 1 taken together, we find no obvious impact of policy on the rate of non-local trackers in each country. In fact, there is a weak negative trend: more permissive countries have *fewer* non-local trackers. Thus adherence with data localization might be more impacted by other constraints (such as availability of nearby data centers). Note that, while Table 1 summarizes national policies concerning cross-border data flows, it does not capture legal compliance or violations by the content providers in our study. Additional information would be required to establish such violations, and the type of information would vary by regulatory regime. In some cases, establishing a violation would require knowledge of the data transferred, for example, personal or non-personal; while in others it would require context about the legal bases for such transfers, such as consent mechanisms, contractual necessity, adequacy decisions.

Data localization requirements may appear attractive to regulators because they can be built on existing principles of territorial sovereignty in each country's legislation. However, content providers have infrastructure that is centralized in a handful of countries and may be unwilling to invest in additional nodes in all countries that mandate data localization, especially if they anticipate smaller revenue potential. Thus, only countries where large tech companies have substantial infrastructure already in place are likely to be able to enforce data localization requirements.

Geographical Distribution and Corporate Control: We conducted an analysis of the geographical distribution and corporate ownership of non-local trackers across all T_{web}. We observed

[1] Though in Azerbaijan, since prior consent of the subject is required, it is difficult to see how the companies could claim that they obtained it from our volunteers.
[2] List not yet published
[3] After opt-out period
[4] Excluding mainland China

Table 1: Data localization policy types [55] by decreasing strictness. **CS**: consent of subject; **PA**: Prior government approval or registration; **AC**: transfers allowed to pre-approved countries; **TA**: transfers allowed if domestic or comparable protections are provided abroad; **NR**: no restrictions

Country	Type	Enacted	Non-Local
Azerbaijan	CS	Yes	74.39%
Algeria	PA	Yes	49.39%
Egypt	PA	Yes	70.41%
Rwanda	PA	Yes	62.30%
Uganda	PA	Yes	75.45%
Argentina	AC	Yes	61.48%
Russia	AC	Yes	8.00%
Sri Lanka	AC	Yes	9.43%
Thailand	AC	No	59.05%
UAE	AC[2]	Yes	33.50%
UK	AC	Yes	38.65%
Australia	TA	Yes	7.06%
Canada	TA	Yes	0.00%
India	TA	No	1.06%
Japan	TA[3]	Yes	22.71%
Jordan	TA	Yes	54.37%
New Zealand	TA	Yes	83.50%
Pakistan	TA	No	65.73%
Qatar	TA	Yes	73.19%
Saudi Arabia	TA	Yes	71.43%
Taiwan	TA[4]	Yes	7.63%
USA	TA	Yes	0.00%
Lebanon	NR	Yes	20.24%

that the servers contacted by these non-local trackers are often in Europe. However, the ownership and the governance of the organizations that operate these tracking mechanisms are heavily dominated by US-based corporations. For instance, all five of the largest tracking organizations in Fig. 8 (Google, Twitter, Facebook, Amazon and Yahoo) are based in the US, as are a substantial portion of the smaller trackers. Thus, while the US does not directly host these trackers in its territory, the major companies that do are headquartered there. Further, many of these companies' servers are hosted in Western European countries and other traditional US allies. This geographic, political and commercial concentration of tracking companies ensures that the economic benefits of emergent technologies flow to a few companies that are clustered together in the Global North, while the costs–e.g., of potential privacy invasions or major leaks–are distributed around the world.

Limitations: While our study provides valuable insights, it is subject to several limitations. We acknowledge that our study only records trackers on the homepage of websites with no interaction (e.g., no scrolling). This limitation may have influenced our findings, as internal pages can behave differently [8]. Moreover, we did not employ any techniques to evade Selenium-based bot detection by websites, and each website was visited once; prior work has shown that Selenium-driven Chrome browsers can influence the visibility of trackers [23], which may have impacted our results. We recommend that future studies perform multiple runs to mitigate

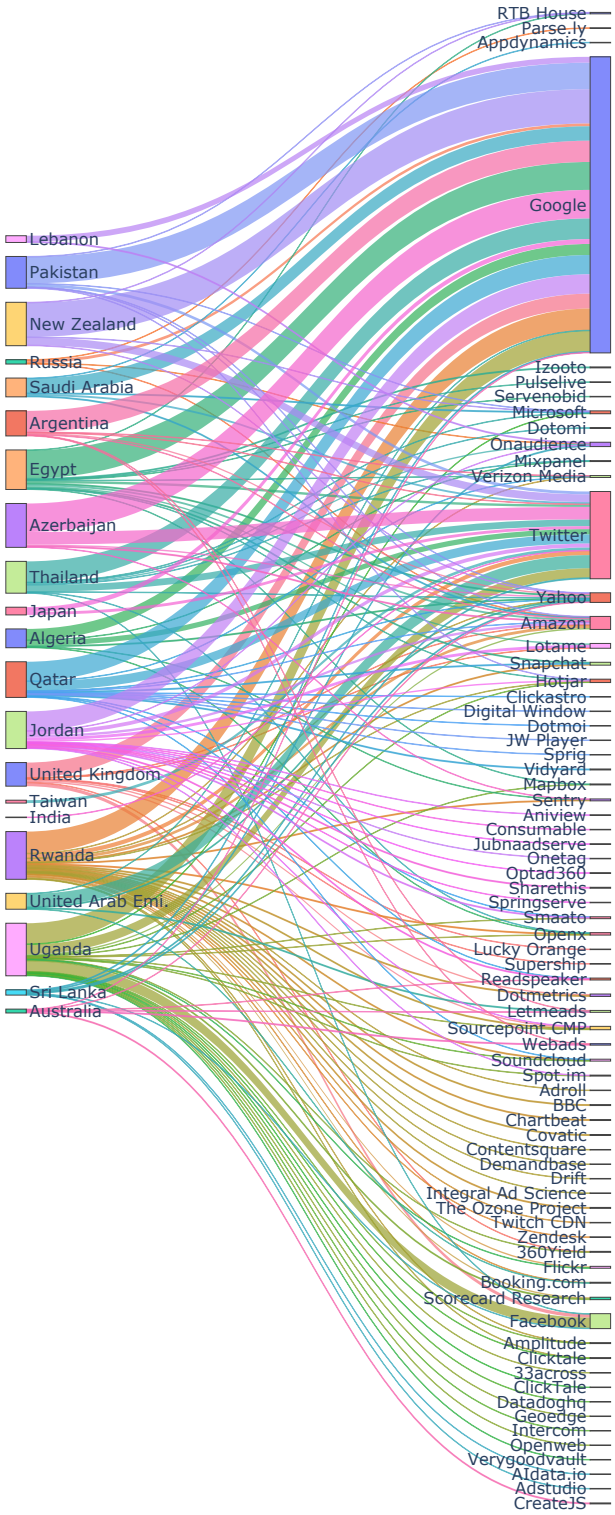


Figure 8: Non-local tracking flows from source countries (left) to organizations (right) operating the tracking domains.

the effects of such variability. We relied on tracker lists and manual inspection to identify trackers and to mitigate list bias. However, as these resources are not exhaustive, and because we excluded trackers that could not be geolocated, our results should be interpreted as a lower bound.

Furthermore, our observations reflect responses from both CDN and edge servers. This reflects where data travels, not a definitive map of where companies store their data, and should be interpreted as a lower bound on the set of countries to which the data is transmitted. Additionally, we relied on RIPE IPmap for geolocation. While it provides useful coverage, it is not always accurate. Although we incorporated other constraints in our study to mitigate this limitation, residual inaccuracies in geolocation may still have influenced our results. Our study is limited both to a point-in-time and to a single ISP in each country. Since we are investigating tracker data flow in many diverse regions where there were few to no such studies previously, we argue that these limitations are reasonable for a first study of this kind.

Recommendations: For *users*, we recommend using privacy-enhancing browser extensions such as EFF's Privacy Badger [43], disable third-party cookies, or use privacy-oriented browsers including Brave, DuckDuckGo or Firefox with hardened settings.

For *operators*, we recommend thorough reviews of embedded third-party scripts and careful consideration before using trackers hosted in offshore locations or jurisdictions when handling sensitive or regulated data. We recognize that accurately geolocating the infrastructure behind these trackers is inherently difficult, which makes it even harder for operators to determine where user data ultimately travels. Conducting such reviews can also impose significant costs, and in cases of noncompliance may result in legal penalties, creating an additional financial burden.

For *tracker organizations*, there is a pressing need to adopt greater transparency regarding data flows. Clearer disclosures about where data is transmitted, stored, and shared would help reduce the opacity that currently prevents website operators and users from understanding the full extent of third-party tracking.

Finally, for *policymakers*, we recommend the execution of technical audits using frameworks that allow for a granular detection of privacy violations, such as the one described in this study and in previous work [48]. These tools allow for empirical quantification of overseas data flows, and are thus a good complement to territorial-based regulation of data sovereignty. We also encourage regulators to lead by example and mandate the disclosure of data collection on governmental websites, particularly when the data is collected by third parties or sent overseas.

8 Conclusion and Future Work

As nations increasingly prioritize privacy regulations and policies, knowledge about the prevalence of foreign web trackers across countries remains limited. Our research provides insights into web tracker data flows based on data collected in 23 diverse countries. We use an integrated approach that involves the distribution of a pre-configured software tool, *Gamma*, to volunteers. We find high variability in the prevalence of foreign trackers across countries. These non-local trackers are hosted in both nearby and remote nations. Europe is a *central hub* for these trackers, which are primarily hosted by US-based companies.

Our recorded data can support further research, such as analyzing local trackers, and privacy and security issues related to tracking domains. The data also serves as a snapshot for longitudinal studies, tracking behavioral changes and regulatory impacts. For example, the Jordanian Data Protection Law [28, 35], effective March 17, 2024, allows our March 16, 2024 recorded data to serve as a baseline for future analysis. Our data also provides a valuable resource for analyzing how the same website can exhibit different behaviors across various countries, particularly in terms of embedded trackers and network requests. For example, Yahoo.com primarily embeds trackers from Yahoo and Google in India and the UK; in contrast, in Australia, Qatar, and the UAE, *Yahoo.com* embeds additional trackers from Demdex (Adobe Audience Manager), Bluekai, and Taboola. This highlights regional adaptations in tracking, which can be further studied using our tool and data.

Acknowledgments

We would like to thank all the volunteers for their dedicated efforts in supporting our research. We also thank the reviewers and our shepherd for their valuable feedback. Finally, we thank Prof. Sam King (UC Davis) for providing comments that helped improve this work. This work was partly funded by the National Science Foundation (NSF) under Award Nos. CNS 2027208 and CNS 2402963. Author Gamero-Garrido was supported in part by the Ford Foundation Postdoctoral Fellowship and the Northeastern Future Faculty Fellowship.

References

[1] ipwhois.io: IP Geolocation API Fast response and accurate data. https://ipwhois.io. Accessed 2025-09-21.
[2] Gamma: Web measurement toolkit. https://github.com/such-in/gamma, 2025. Accessed 2025-09-21.
[3] Zainul Abi Din, Panagiotis Tigas, Samuel T King, and Benjamin Livshits. Percival : Making in-browser perceptual ad blocking practical with deep learning. In *2020 USENIX Annual Technical Conference (USENIX ATC 20)*, pages 387–400, 2020.
[4] ahrefs. Top websites. https://ahrefs.com/top. Accessed 2024-04-04.
[5] Mshabab Alrizah, Sencun Zhu, Xinyu Xing, and Gang Wang. Errors, misunderstandings, and attacks: Analyzing the crowdsourcing process of ad-blocking systems. In *Proceedings of the Internet Measurement Conference*, pages 230–244, 2019.
[6] ANPDP. National authority for the protection of personal data. https://www.joradp.dz/FTP/JO-FRANCAIS/2018/F2018034.pdf. Accessed 2024-05-05.
[7] Uyên P. Lê Anupam Chander. Data nationalism. https://scholarlycommons.law.emory.edu/elj/vol64/iss3/2/. Accessed 2024-04-04.
[8] Waqar Aqeel, Balakrishnan Chandrasekaran, Anja Feldmann, and Bruce M Maggs. On landing and internal web pages: The strange case of jekyll and hyde in web performance measurement. In *Proceedings of the ACM Internet Measurement Conference*, pages 680–695, 2020.
[9] Argentina.gob.ar. Personal data protection. https://www.argentina.gob.ar/aaip/datospersonales. Accessed 2024-05-05.
[10] AWS. Amazon cloudfront key features. https://aws.amazon.com/cloudfront/features/?p=ugi&l=emea&whats-new-cloudfront.sort-by=item.additionalFields.postDateTime&whats-new-cloudfront.sort-order=desc. Accessed 2024-05-05.
[11] AWS. What is a cdn (content delivery network)? https://aws.amazon.com/what-is/cdn/. Accessed 2024-08-04.
[12] Muhammad Ahmad Bashir, Sajjad Arshad, Engin Kirda, William Robertson, and Christo Wilson. How tracking companies circumvented ad blockers using websockets. In *Proceedings of the Internet Measurement Conference 2018*, pages 471–477, 2018.
[13] Nataliia Bielova. Web tracking technologies and protection mechanisms. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications*

Security, pages 2607–2609, 2017.
[14] Reuben Binns et al. Tracking on the web, mobile and the internet of things. *Foundations and Trends® in Web Science*, 8(1–2):1–113, 2022.
[15] Bright Data. Bright data: Limitless web data infrastructure for AI & BI. https://brightdata.com. Accessed 2025-05-10.
[16] Duc Bui, Brian Tang, and Kang G Shin. Do opt-outs really opt me out? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 425–439, 2022.
[17] Michael Butkiewicz, Harsha V Madhyastha, and Vyas Sekar. Understanding website complexity: measurements, metrics, and implications. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 313–328, 2011.
[18] CAIDA. Archipelago (ark): Caida's active measurement infrastructure serving the network research. https://www.caida.org/projects/ark/locations/. Accessed 2024-08-04.
[19] CAIDA. Mapping autonomous systems to organizations. https://www.caida.org/archive/as2org/. Accessed 2024-05-05.
[20] Massimo Candela, Enrico Gregori, Valerio Luconi, and Alessio Vecchio. Dissecting the speed-of-internet of middle east. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 720–725. IEEE, 2019.
[21] Massimo Candela, Enrico Gregori, Valerio Luconi, and Alessio Vecchio. Using ripe atlas for geolocating ip infrastructure. *IEEE Access*, 7:48816–48829, 2019.
[22] Massimo Candela, Valerio Luconi, and Alessio Vecchio. A worldwide study on the geographic locality of internet routes. *Computer Networks*, 201:108555, 2021.
[23] Darion Cassel, Su-Chin Lin, Alessio Buraggina, William Wang, Andrew Zhang, Lujo Bauer, Hsu-Chun Hsiao, Limin Jia, and Timothy Libert. Omnicrawl: Comprehensive measurement of web tracking with real desktop and mobile browsers. *Proceedings on Privacy Enhancing Technologies*, 2022.
[24] Quan Chen, Peter Snyder, Ben Livshits, and Alexandros Kapravelos. Detecting filter list evasion with event-loop-turn granularity javascript signatures. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 1715–1729. IEEE, 2021.
[25] Jinchun Choi, Mohammed Abuhamad, Ahmed Abusnaina, Afsah Anwar, Sultan Alshamrani, Jeman Park, Daehun Nyang, and David Mohaisen. Understanding the proxy ecosystem: A comparative analysis of residential and open proxies on the internet. *IEEE Access*, 8:111368–111380, 2020.
[26] Cloudflare. What is a content delivery network (cdn)? | how do cdns work? https://www.cloudflare.com/learning/cdn/what-is-a-cdn/. Accessed 2024-08-04.
[27] Cloudwards. Netflix vpn not working? https://www.cloudwards.net/how-to-beat-the-netflix-vpn-ban/. Accessed 2024-04-04.
[28] Clydeco. Jordan issues first personal data protection law. https://www.clydeco.com/en/insights/2023/10/jordan-issues-first-personal-data-protection-law. Accessed 2024-04-04.
[29] CNET. Geo-blocking explained: What to know and how you can get around it. https://www.cnet.com/tech/services-and-softwar/what-is-geo-blocking-and-how-you-can-get-around-it/. Accessed 2024-04-04.
[30] Comparitech. Where are vpns legal and where are they banned? https://www.comparitech.com/vpn/where-are-vpns-legal-banned/. Accessed 2024-04-04.
[31] Miguel Cozar, David Rodriguez, Jose M Del Alamo, and Danny Guaman. Reliability of ip geolocation services for assessing the compliance of international data transfers. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 181–185. IEEE, 2022.
[32] cyberhost. 10 countries where vpns are illegal. find out who banned vpns and why. https://www.cyberghostvpn.com/en_US/privacyhub/countries-banning-vpn/. Accessed 2024-04-04.
[33] Savino Dambra, Iskander Sanchez-Rola, Leyla Bilge, and Davide Balzarotti. When sally met trackers: Web tracking from the users' perspective. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2189–2206, 2022.
[34] Datacenters.com. Countries with data centers. https://www.datacenters.com/locations/countries. Accessed 2024-05-05.
[35] Dataguidance. Jordan: An overview of the data protection law for 2023. https://www.dataguidance.com/opinion/jordan-overview-data-protection-law-2023. Accessed 2024-04-04.
[36] Dataguidance. Malaysia - data protection overview. https://www.dataguidance.com/notes/malaysia-data-protection-overview. Accessed 2024-04-04.
[37] DB-IP. The db-ip database. https://db-ip.com/. Accessed 2024-04-04.
[38] Ben Du, Massimo Candela, Bradley Huffaker, Alex C Snoeren, and KC Claffy. Ripe ipmap active geolocation: Mechanism and performance evaluation. *ACM SIGCOMM Computer Communication Review*, 50(2):3–10, 2020.
[39] Easylist. The goal of easylist is to block ads on english and international sites. https://easylist.to/pages/policy.html.
[40] EasyList. A primary filter list that removes most adverts from international webpages. https://easylist.to/easylist/easylist.txt. Accessed 2024-04-04.
[41] EasyPrivacy. An optional supplementary filter list that completely removes all forms of tracking from the internet. https://easylist.to/easylist/easyprivacy.txt. Accessed 2024-04-04.
[42] K Egevang. The ip network address translator (nat). Technical report, IETF RFC 1631, 1994.
[43] Electronic Frontier Foundation. Privacy badger. https://privacybadger.org, 2024. Accessed 2025-05-14.
[44] Digital Element. Netacuity industry-standard geolocation - digital element. https://www.digitalelement.com/geolocation/. Accessed 2024-04-04.
[45] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 1388–1401, 2016.
[46] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. The rise of panopticons: Examining region-specific third-party web tracking. In *Traffic Monitoring and Analysis: 6th International Workshop, TMA 2014, London, UK, April 14, 2014. Proceedings 6*, pages 104–114. Springer, 2014.
[47] Forbes. Are vpns legal? the worldwide guide. https://www.forbes.com/advisor/business/are-vpns-legal/. Accessed 2024-04-04.
[48] Alexander Gamero-Garrido, Kicho Yu, Sumukh Vasisht Shankar, Sachin Kumar Singh, Sindhya Balasubramanian, Alexander Wilcox, and David Choffnes. Empirically measuring data localization in the eu. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2025. To appear. Available at: https://arxiv.org/abs/2504.09019.
[49] Oliver Gasser, Quirin Scheitle, Pawel Foremski, Qasim Lone, Maciej Korczyński, Stephen D Strowes, Luuk Hendriks, and Georg Carle. Clusters in the expanse: Understanding and unbiasing ipv6 hitlists. In *Proceedings of the Internet Measurement Conference 2018*, pages 364–378, 2018.
[50] Manaf Gharaibeh, Anant Shah, Bradley Huffaker, Han Zhang, Roya Ensafi, and Christos Papadopoulos. A look at router geolocation in public and commercial databases. In *Proceedings of the 2017 Internet Measurement Conference*, pages 463–469, 2017.
[51] Github. Adblock india. https://easylist-downloads.adblockplus.org/indianlist.txt. Accessed 2024-08-04.
[52] Github. Adblock srilanka. https://github.com/miyurusankalpa/adblock-list-srilanka. Accessed 2024-08-04.
[53] Google Public DNS. Edns client subnet (ecs) guidelines. https://developers.google.com/speed/public-dns/docs/ecs. Accessed 2025-09-23.
[54] Matthias Gotze, Srdjan Matic, Costas Iordanou, Georgios Smaragdakis, and Nikolaos Laoutaris. Measuring web cookies in governmental websites. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 44–54, 2022.
[55] Data Guidance. Data protection. https://www.dataguidance.com/. Accessed 2024-05-05.
[56] Data Guidance. India - data protection. https://www.dataguidance.com/jurisdiction/india. Accessed 2024-05-05.
[57] Data Guidance. Understanding india's new data protection law. https://carnegieindia.org/research/2023/10/understanding-indias-new-data-protection-law. Accessed 2024-05-05.
[58] John Hawley. Geodns-geographically-aware, protocol-agnostic load balancing at the dns level. In *Proceedings of the linux symposium*, pages 123–130. Citeseer, 2009.
[59] Bradley Huffaker, Marina Fomenkov, and KC Claffy. Geocompare: a comparison of public and commercial geolocation databases. *Proc. NMMC*, pages 1–12, 2011.
[60] ICO. Data minimisation. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/8-data-minimisation/. Accessed 2024-08-04.
[61] imewecable. Imewe (india-middle east-western europe) submarine cable. https://imewecable.com/aboutus.jp. Accessed 2024-04-04.
[62] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. Tracing cross border web tracking. In *Proceedings of the internet measurement conference 2018*, pages 329–342, 2018.
[63] IPinfo.io. The trusted source for ip address data. https://ipinfo.io/. Accessed 2024-04-04.
[64] RIPE IPmap. Ripe ipmap is the ripe ncc's tool for mapping core internet infrastructure. https://ipmap.ripe.net/. Accessed 2024-04-04.
[65] Umar Iqbal, Zubair Shafiq, and Zhiyun Qian. The ad wars: retrospective measurement and analysis of anti-adblock filter lists. In *Proceedings of the 2017 Internet Measurement Conference*, pages 171–183, 2017.
[66] Bernardus Jansen, Natalia Kadenko, Dennis Broeders, Michel van Eeten, Kevin Borgolte, and Tobias Fiebig. Pushing boundaries: An empirical view on the digital sovereignty of six governments in the midst of geopolitical tensions. *Government Information Quarterly*, 40(4):101862, 2023.
[67] Ethan Katz-Bassett, John P John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. Towards ip geolocation using delay and topology measurements. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 71–84, 2006.
[68] kenyanwallstreet. Amazon set to launch an aws local zone in kenya. https://kenyanwallstreet.com/amazon-set-to-launch-an-aws-local-zone-in-kenya/. Accessed 2024-05-05.
[69] Mohammad Taha Khan, Joe DeBlasio, Geoffrey M Voelker, Alex C Snoeren, Chris Kanich, and Narseo Vallina-Rodriguez. An empirical analysis of the commercial vpn ecosystem. In *Proceedings of the Internet Measurement Conference 2018*, pages 443–456, 2018.

[70] Sakwa Kombo. How kenyans connect to the internet. https://techweez.com/2024/04/05/how-kenya-access-the-internet/. Accessed 2024-05-05.

[71] Rashna Kumar, Sana Asif, Elise Lee, and Fabian E Bustamante. Each at its own pace: Third-party dependency and centralization around the world. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1):1–29, 2023.

[72] Rashna Kumar, Esteban Carisimo, Lukas De Angelis Riva, Mauricio Buzzone, Fabián E Bustamante, Ihsan Ayyub Qazi, and Mariano G Beiró. Of choices and control-a comparative analysis of government hosting. In *Proceedings of the 2024 ACM on Internet Measurement Conference*, pages 462–479, 2024.

[73] European Union Law. General data protection regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016. Accessed 2024-05-05.

[74] Ada Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *25th USENIX Security Symposium (USENIX Security 16)*, 2016.

[75] Tai-Ching Li, Huy Hang, Michalis Faloutsos, and Petros Efstathopoulos. Trackadvisor: Taking back browsing privacy from third-party trackers. In *Passive and Active Measurement: 16th International Conference, PAM 2015, New York, NY, USA, March 19-20, 2015, Proceedings 16*, pages 277–289. Springer, 2015.

[76] Ioana Livadariu, Thomas Dreibholz, Anas Saeed Al-Selwi, Haakon Bryhni, Olav Lysne, Steinar Bjørnstad, and Ahmed Elmokashfi. On the accuracy of country-level ip geolocation. In *Proceedings of the applied networking research workshop*, pages 67–73, 2020.

[77] Matthew Luckie, Bradley Huffaker, Alexander Marder, Zachary Bischof, Marianne Fletcher, and K Claffy. Learning to extract geographic information from internet router hostnames. In *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies*, pages 440–453, 2021.

[78] MaxMind. Industry leading ip geolocation. https://www.maxmind.com/en/home. Accessed 2024-04-04.

[79] MeasurementLab. Measurement lab: Measure the internet, save the data, and make it universally accessible and useful. https://www.measurementlab.net/status/. Accessed 2024-08-04.

[80] Xianghang Mi, Xuan Feng, Xiaojing Liao, Baojun Liu, XiaoFeng Wang, Feng Qian, Zhou Li, Sumayah Alrwais, Limin Sun, and Ying Liu. Resident evil: Understanding residential ip proxy as a dark service. In *2019 IEEE symposium on security and privacy (SP)*, pages 1185–1201. IEEE, 2019.

[81] Netflix. Netflix error m7111-5059. https://help.netflix.com/en/node/53047. Accessed 2024-04-04.

[82] Netflix. Netflix says, 'you seem to be using a vpn or proxy.'. https://help.netflix.com/en/node/277. Accessed 2024-04-04.

[83] Submarine Cable Network. Bharat lanka cable system overview. https://www.submarinenetworks.com/en/systems/intra-asia/blcs/bharat-lanka-cable-system. Accessed 2024-04-04.

[84] Submarine Networks. There are now 6 submarine cables landing in kenya. https://www.submarinenetworks.com/en/stations/africa/kenya. Accessed 2024-05-05.

[85] Irish Tech News. Why do websites block vpns? https://irishtechnews.ie/why-do-websites-block-vpns/. Accessed 2024-04-04.

[86] NLNOG. Nlnog ring. https://ring.nlnog.net/nodes/. Accessed 2024-08-04.

[87] Nmap. Nmap: the network mapper - free security scanner. https://nmap.org/. Accessed 2024-08-04.

[88] Daiyuu Nobori and Yasushi Shinjo. {VPN} gate: A {Volunteer-Organized} public {VPN} relay system with blocking resistance for bypassing government censorship firewalls. In *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*, pages 229–241, 2014.

[89] Times of India. Yahoo shuts down news sites in india. https://timesofindia.indiatimes.com/business/india-business/yahoo-shuts-down-news-sites-in-india/articleshow/85651433.cms. Accessed 2024-04-04.

[90] perfSONAR. perfsonar is the performance service-oriented network monitoring architecture, a network measurement toolkit. https://stats.perfsonar.net/d/db0e6ecb-3fe3-46b3-9745-edf6a209e7cf/. Accessed 2024-08-04.

[91] PlanetLab. Planetlab eu testbed. https://www.planet-lab.eu/. Accessed 2024-08-04.

[92] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. *arXiv preprint arXiv:1806.01156*, 2018.

[93] Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. Ip geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review*, 41(2):53–56, 2011.

[94] Reethika Ramesh, Philipp Winter, Sam Korman, and Roya Ensafi. Calculatency leveraging cross-layer network latency measurements to detect proxy-enabled abuse. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2263–2280, 2024.

[95] Kimberly Ruth, Aurore Fass, Jonathan Azose, Mark Pearson, Emma Thomas, Caitlin Sadowski, and Zakir Durumeric. A world wide view of browsing the world wide web. In *Proceedings of the 22nd ACM Internet Measurement Conference*, pages 317–336, 2022.

[96] Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. Toppling top lists: Evaluating the accuracy of popular website lists. In *Proceedings of the 22nd ACM Internet Measurement Conference*, pages 374–387, 2022.

[97] Nayanamana Samarasinghe, Aashish Adhikari, Mohammad Mannan, and Amr Youssef. Et tu, brute? privacy analysis of government websites and mobile apps. In *Proceedings of the ACM Web Conference 2022*, pages 564–575, 2022.

[98] Nayanamana Samarasinghe and Mohammad Mannan. Towards a global perspective on web tracking. *Computers & Security*, 87:101569, 2019.

[99] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. Can i opt out yet? gdpr and the global illusion of cookie control. In *Proceedings of the 2019 ACM Asia conference on computer and communications security*, pages 340–351, 2019.

[100] Iskander Sanchez-Rola and Igor Santos. Knockin'on trackers' door: Large-scale automatic analysis of web tracking. In *Detection of Intrusions and Malware, and Vulnerability Assessment: 15th International Conference, DIMVA 2018, Saclay, France, June 28–29, 2018, Proceedings 15*, pages 281–302. Springer, 2018.

[101] Scapy. Scapy is a powerful interactive packet manipulation library written in python. https://scapy.net/. Accessed 2024-08-04.

[102] Github Scapy. Scapy is a powerful python-based interactive packet manipulation program and library. https://github.com/secdev/scapy. Accessed 2024-08-04.

[103] Tomer Schwartz, Ofir Manor, and Andikan Otung. Snitch: Leveraging ip geolocation for active vpn detection. In *Proceedings of Workshop on Measurements, Attacks, and Defenses for the Web (MadWeb)*. Network and Distributed System Security Symposium (NDSS), 2025.

[104] Semrush. Top websites wcross the web. https://www.semrush.com/trending-websites/. Accessed 2024-04-04.

[105] Yuval Shavitt and Noa Zilberman. A geolocation databases study. *IEEE Journal on Selected Areas in Communications*, 29(10):2044–2056, 2011.

[106] Similarweb. Top website list. https://www.similarweb.com/top-websites/. Accessed 2024-04-04.

[107] Alexander Sjösten, Peter Snyder, Antonio Pastor, Panagiotis Papadopoulos, and Benjamin Livshits. Filter list generation for underserved regions. In *Proceedings of The Web Conference 2020*, pages 1682–1692, 2020.

[108] Peter Snyder, Antoine Vastel, and Ben Livshits. Who filters the filters: Understanding the growth, usefulness and efficiency of crowdsourced ad blocking. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(2):1–24, 2020.

[109] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. Flash cookies and privacy. In *2010 AAAI Spring Symposium Series*, 2010.

[110] Jannick Sørensen and Sokol Kosta. Before and after gdpr: The changes in third party presence at public and private european websites. In *The World Wide Web Conference*, pages 1590–1600, 2019.

[111] Statista.com. Leading countries by number of data centers as of march 2024. https://www.statista.com/statistics/1228433/data-centers-worldwide-by-country/. Accessed 2024-05-05.

[112] Telegeography. Cloud infrastructure map. https://www.cloudinfrastructuremap.com/. Accessed 2024-04-04.

[113] Telegeography. Where are the world's cloud data centers. https://blog.telegeography.com/where-are-the-worlds-cloud-data-centers-and-who-is-using-them. Accessed 2024-04-04.

[114] the star.co.ke. Amazon web services development centre launched in kenya. https://aws.amazon.com/about-aws/global-infrastructure/localzones/locations/. Accessed 2024-05-05.

[115] Business Today. Yahoo shuts down news operations in india; yahoo mail continues to operate. https://www.businesstoday.in/latest/corporate/story/yahoo-shuts-down-news-operations-in-india-yahoo-mail-continues-to-operate-305215-2021-08-26. Accessed 2024-04-04.

[116] Tom'sguide. Countries with the strictest vpn laws. https://www.tomsguide.com/features/8-countries-with-the-strictest-vpn-laws. Accessed 2024-04-04.

[117] Tranco. Tranco list. https://tranco-list.eu/. Accessed 2024-04-04.

[118] Lonneke Van der Velden. The third party diary–tracking the trackers on dutch governmental websites. *NECSUS. European Journal of Media Studies*, 3(1):195–217, 2014.

[119] Verizon. Ip latency statistics. https://www.verizon.com/business/terms/latency/. Accessed 2024-04-04.

[120] VPN.com. Which countries block vpns and why in 2024? https://www.vpn.com/guide/which-countries-block-vpn/. Accessed 2024-04-04.

[121] Zachary Weinberg, Shinyoung Cho, Nicolas Christin, Vyas Sekar, and Phillipa Gill. How to catch when proxies lie: Verifying the physical locations of network proxies with active geolocation. In *Proceedings of the Internet Measurement Conference 2018*, pages 203–217, 2018.

[122] Dirk Wetter. testssl.sh is a free command line tool which checks a server's service on any port for the support of tls/ssl ciphers. https://testssl.sh. Accessed 2024-08-04.

[123] Whotracks.me. Trackers. https://whotracks.me/trackers.html. Accessed 2024-04-04.

[124] Whotracks.me. Gdpr - what happened? https://whotracks.me/blog/gdpr-what-happened.html, 2018. Accessed 2024-05-05.
[125] Craig E Wills and Doruk C Uzunoglu. What ad blockers are (and are not) doing. In *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, pages 72–77. IEEE, 2016.
[126] WonderNetwork. Global ping statistics. https://wondernetwork.com/pings. Accessed 2024-04-04.
[127] Yahoo. Yahoo stops publication of content in india citing new fdi regulation. https://ca.movies.yahoo.com/movies/yahoo-stops-publication-content-india-044949199.html. Accessed 2024-04-04.
[128] Yext. People overwhelmingly trust government websites — but can the public sector deliver? https://www.yext.com/blog/2021/12/people-overwhelmingly-trust-government-websites-but-can-the-public-sector-deliver. Accessed 2024-04-04.
[129] Ahsan Zafar and Anupam Das. Comparative privacy analysis of mobile browsers. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, pages 3–14, 2023.
[130] Alexander R Zheutlin, Joshua D Niforatos, and Jeremy B Sussman. Data-tracking on government, non-profit, and commercial health-related websites. *Journal of general internal medicine*, pages 1–3, 2021.

Appendix

A Non-Local Tracking Domain Distribution

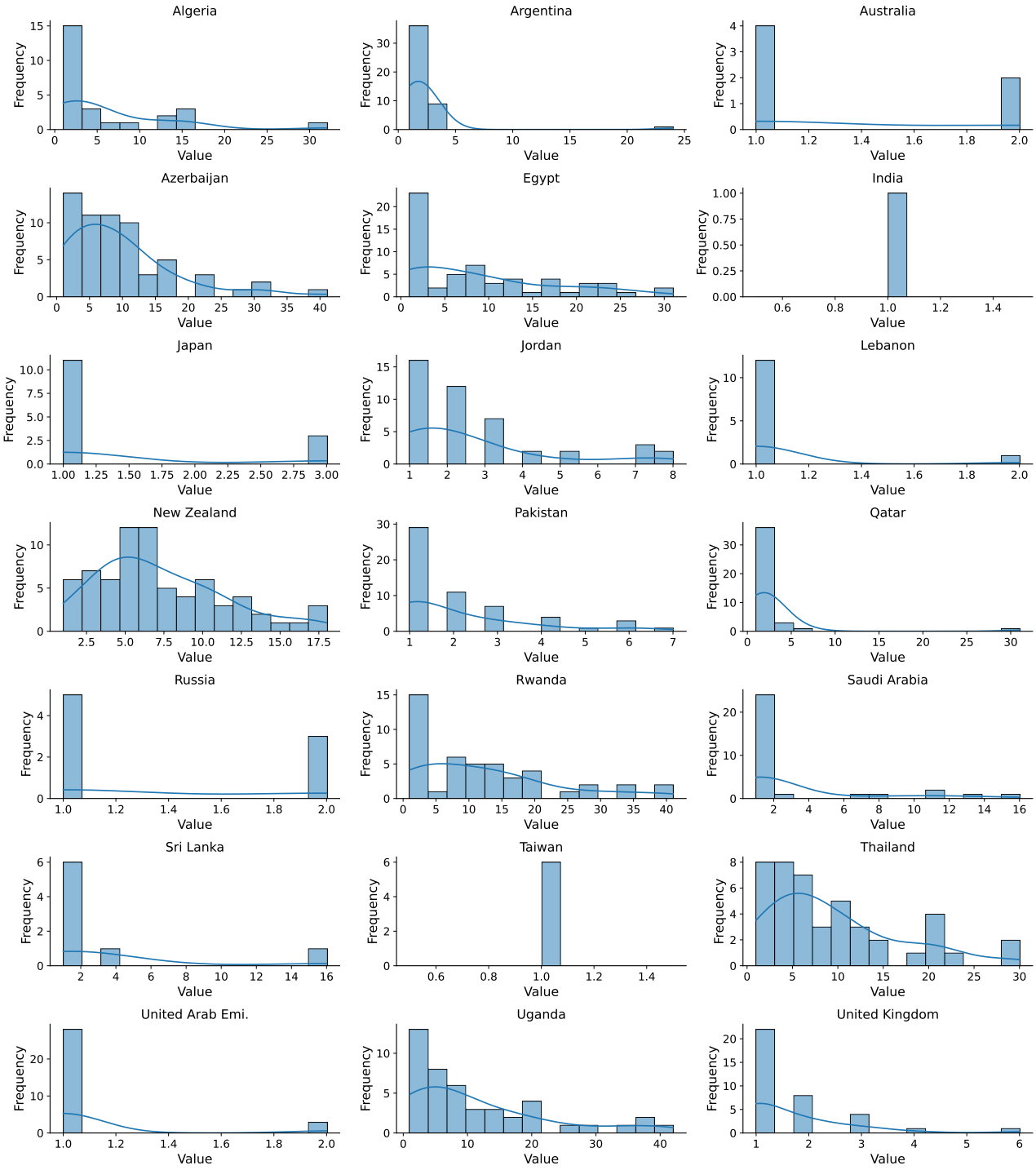


Figure 9: Frequency of non-local tracking domains across various websites in various countries.