Empirically Measuring Data Localization in the EU

Alexander Gamero-Garrido
University of California, Davis

Kicho Yu
University of Southern California

Sumukh Vasisht Shankar
Yale University

Sachin Kumar Singh
University of Utah

Sindhya Balasubramanian
Northeastern University

Alexander Wilcox
Northeastern University

David Choffnes
Northeastern University

ABSTRACT

EU data localization regulations limit data transfers to non-EU countries with the GDPR. However, BGP, DNS and other Internet protocols were not designed to enforce jurisdictional constraints, so implementing data localization is challenging. Despite initial research on the topic, little is known about if or how companies currently operate their server infrastructure to comply with the regulations. We close this knowledge gap by empirically measuring the extent to which servers and routers that process EU requests are located outside of the EU (and a handful of "adequate" non-EU countries). The key challenge is that both browser measurements (to infer relevant endpoints) and data-plane measurements (to infer relevant IP addresses) are needed, but no large-scale public infrastructure allows both. We build a novel methodology that combines BrightData (browser) and RIPE Atlas (data-plane) probes, with joint measurements from over 1,000 networks in 20 EU countries. We find that, on average, 2.2% of servers serving users in each EU country are located in non-adequate destination countries (1.4% of known trackers). Our findings suggest that data localization policies are largely being followed by content providers, though there are exceptions.

KEYWORDS

Data Localization, GDPR, Internet Measurement

1 INTRODUCTION

Regulators around the world have taken various approaches at mitigating potential harms resulting from unfettered collection of user data at a large scale by Internet companies, usually for the purpose of targeted advertising. Data localization—broadly defined as the principle that requires user data to be stored in the jurisdiction where the user is physically located—is one such regulatory approach and a key component of the EU's privacy landmark regulation enacted in 2018: The General Data Protection Regulation (GDPR). A central intent behind (GDPR) data localization is to prevent transfers of user data to jurisdictions with weaker privacy protections, as such transfers expose (EU) users to unlawful commercial surveillance of user data by both commercial advertisers and foreign government agencies, as was argued in the landmark *Schrems* decisions [46, 58].

In this paper, we explore the question of whether content providers serving users in Europe comply with data localization regulations by locating their servers in the EU (or a small number of third countries that the EU regulators have approved data transfers to [13]). In answering this question, we address a key technical challenge. Since data localization is based on physical boundaries, three types of information are simultaneously needed to audit compliance with data localization, and no existing platform nor methodology can procure all three: (*i*) Which domains are most frequently visited by European users, including many additional domains fetched automatically by user-loaded websites. (*ii*) The location in the network of all relevant domains, that is, their IP addresses. (*iii*) The physical location of the servers from which these domains are loaded. While (*iii*) is the key information needed to audit data localization practices, in practical terms it is difficult to obtain it directly from (*i*) domain names; rather, it is easier (though still challenging [33]) to obtain the physical location of IP addresses, a necessary resource to connect to the Internet an actual server referenced by each domain.

Our auditing framework on data localization compliance is built with three components, each obtaining the necessary information described above. First, we identify all popular domains in each EU country using queries from BrightData [8], a proxy service that allows us to run a headless browser and therefore fetch complete web pages from EU-based devices. Second, we launch active measurements from RIPE Atlas [3] to identify the IP addresses serving each domain. Third, we use RIPE IPMap, [33]) a database that translates IP addresses to a physical location (IP geolocation), to identify IP addresses that are potentially located in countries outside the EU. We confirm a subset of these as likely GDPR violations using active measurements and speed-of-light constraints. We validate our method using servers with known locations, finding no false positives.

We find that the vast majority of servers responding to requests from EU users are located in the EU or adequate third countries. Neighbors of the EU, Russia and Turkey, are the most commonly observed destinations in non-adequate countries, besides the US. These servers primarily serve requests in Finland and Romania, respectively. We find smaller numbers of servers and known trackers in other non-adequate countries further away from the EU, primarily in Asia. News websites are the leading cause of these potential GDPR violations, as they are the most common type of site that loads trackers in non-adequate countries. We find evidence of tracking activity taking place in a non-adequate country

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit https://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.
Proceedings on Privacy Enhancing Technologies YYYY(X), 1–15
© YYYY Copyright held by the owner/author(s).
https://doi.org/XXXXXXX.XXXXXXX

using cookies as well. We also find significant differences across EU regions: Southern and Eastern EU countries are more likely to be served content by known trackers located in non-adequate countries when compared to Western and Northern EU countries.

Combined, these findings suggest that Internet companies are likely inclined to host content in close proximity of users, which leads to by-default compliance with data localization regulations; however, there are exceptions where rigid technical constraints may force Internet operators to serve EU users from non-compliant jurisdictions. In these circumstances, the imposition of physical constraints on the Internet infrastructure is likely to be more challenging than currently recognized by both Internet operators and regulators alike, which is especially relevant as proposed regulations on data localization emerge in more countries [19, 39].

2 POLICY BACKGROUND

Data transfers to jurisdictions with weaker privacy protections expose users in the initial jurisdiction to potential harms such as leaks of personal information to foreign governments or advertisers. Despite these potential harms, little is known about whether or how Internet companies that serve EU users operate their infrastructure to comply with this provision of the GDPR (Chapter V).

Specifically, the GDPR in Article 45 mentions that "A transfer of personal data to a third country or an international organisation may take place where the [European] Commission has decided that the third country, a territory or one or more specified sectors within that third country, or the international organisation in question ensures an adequate level of protection." [17] Generally speaking, the article refers to several principles that must be followed in such countries receiving data transfers from EU persons, such as "elements like rule of law, respect for human rights and fundamental freedoms, as well as whether or not data subjects' rights are effective and enforceable, the existence and effective functioning of an independent data protection authority in the non-EEA country and the international commitments the country or international organisation has entered into." [7] These adequacy decisions are published on a European Commission (EC) website [13] and the decisions must be reviewed and renewed periodically by law. At the time of writing, the EC has determined that transfers between the EU and other European Economic Area countries (Norway, Liechtenstein, and Iceland) are treated as intra-EU transfers without needing any further safeguard. Also, according to that same website, the EC "has so far recognised Andorra, Argentina, Canada (commercial organisations), Faroe Islands, Guernsey, Israel, Isle of Man, Japan, Jersey, New Zealand, Republic of Korea, Switzerland, the United Kingdom under the GDPR and the LED, the United States (commercial organisations participating in the EU-US Data Privacy Framework) and Uruguay as providing adequate protection."

Any countries not covered by the above list are potentially inadequate, and any international data transfers from the EU to them would violate Article 45 if that country is deemed inadequate by EU courts, *e.g.*, the Court of Justice of the European Union (CJEU) or the EC. Even countries previously determined to be adequate may face a reversal of that decision. For example, in 2020, the CJEU determined in *Schrems II* [46] that "As a consequence of such a degree of interference with the fundamental rights of persons whose data are transferred to [the United States, where national security laws limit data protections], the Court declared the Privacy Shield adequacy Decision invalid." [6] We note that data transfers to the US were not broadly authorized when we collected our data in Aug-Sept 2022, as the EU-US Data Privacy Framework was not adopted until December of that year [14]. The EU-US DPF was adopted in response to the *Schrems II* decision and determined adequate by the EC in July 2023. [16]

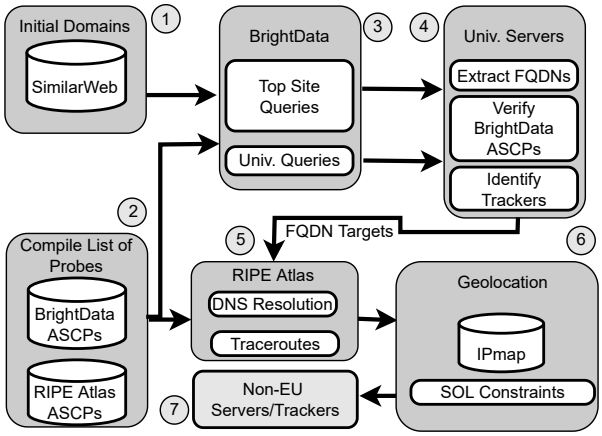


Figure 1: Process diagram summarizing our methodology. Steps numbered (1)&(2) are data sources. Steps (3)&(4) are browser-based measurements. Steps (5)&(6) are data-plane measurements. Step (7) refers to rDNS filtering and the final output. All steps are referenced in our method sections.

Given regular instances of noncompliance resulting in EU-imposed penalties on Internet companies [20, 57], it is clear that content providers are not universally complying with these GDPR provisions. Besides these anecdotal instances, and studies with limited relevance (see § 11), there is currently no EU-wide study of whether data localization principles are being complied with in practice by content providers.

Indeed, privacy advocates have primarily relied on the *possibility* that data may be transferred to a jurisdiction with weaker privacy protections, especially transatlantic transfers which might result in exposure to US government surveillance. While these claims have successfully moved the needle on public policy (holding large tech companies such as Google accountable for transatlantic data transfers [26]) through the European courts, they are based on analyses of the tech companies' own privacy policies and manual, surface-level analysis of HTML [44]; thus, these methodologies are neither aimed at nor adequate to audit compliance of data localization by content providers at EU scale.

In this paper, we focus on data transfers to countries not determined to be adequate at the time of measurement (see definition of adequate above). We refer to such countries as *non-adequate*. Thus, our study identifies *potential* GDPR violations, that is the existence of a tracking server in a non-adequate country.

3 NETWORK & COUNTRY SAMPLE

Our method ensures that we launch both browser-based and network measurements from the same network and in the same country. (We do not collect any personal information; we include an

ethics section in the appendix). This constraint significantly increases the likelihood that both the DNS resolution, and therefore also the responding server, are identical in both sets of measurements. To this end, we identify overlaps in the measurement infrastructure provided by two platforms: RIPE, the European Internet registrar that hosts RIPE Atlas [50], a large-scale Internet measurement platform with very dense deployment in the EU; and BrightData, a large-scale proxy service [8]. To compute this overlap, we look at the networks present in each EU country that are represented in each platform. We evaluate networks at the granularity of Autonomous Systems (AS), the administrative domain over which a company (network operator) has control over; ASes are the entities explicitly named in entries on the routing system, the Border Gateway Protocol, that rules over how traffic is delivered in the global Internet. We thus look for AS-Country Pairs (ASCPs), or an AS in a country–a single AS can operate in multiple countries–where both platforms host a probe, step (2) of Fig. 1.

While RIPE regularly publishes a list of its active probes [3], including country and AS, BrightData does not provide a list of active networks in each country. To find BrightData's AS-Country Pairs in the EU, we send repeated queries to request a proxy in a specific country over a period of two weeks in the last quarter of 2021. We find that while RIPE has presence in 2,957 ASCPs, BrightData is present in 4,037. The intersection is 1,355 ASCPs, covering 1,318 ASes in 27 countries.

3.1 BrightData Justification

BrightData ensures that traffic comes from 1,000+ residential EU networks, reflecting what EU residents see on their browsers. Other proxy services lack this coverage. Further, RIPE Atlas, while widespread in the EU, lacks a browser to execute JavaScript, limiting its capability; data center-based approaches, meanwhile, are easily detectable and treated differently by content providers. Although BrightData can not load Google domains as initial sites, we observed 25 out of 41 Google domains (see § 4.3) as third parties on non-Google sites, capturing key interactions between browsers and Google properties. We conduct an experiment to validate BrightData's proxy locations in § 9.

4 IDENTIFYING RELEVANT DOMAINS & TRACKERS

In this section, we describe our identification of relevant, popular domains in each EU country, step (1) in Fig. 1. Our code and data is available on GitHub. [25]

4.1 Initial Sample of Top Sites per EU Country

We rely on a list of the top 50 websites in each EU country published by SimilarWeb [55], which has also been used in previous studies [10, 62]. SimilarWeb is the source of initial domains represented in step (1) of Fig. 1. From this list, we exclude 19 adult sites as queries to them are not permitted by BrightData. SimilarWeb has no list of top sites in 7 smaller EU countries, so we exclude them from our sample. We are left with 604 websites in 20 countries.

To obtain a representative sample of popular websites in each EU country, we used SimilarWeb instead of Tranco, [35] as tranco primarily focuses on global rankings. Previous work has shown that

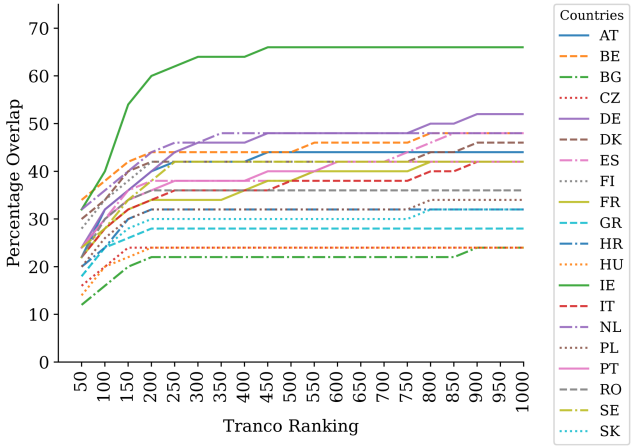


Figure 2: Percentage of overall agreement between Tranco and SimilarWeb lists.

global website lists are skewed toward certain nations and do not reflect regional experiences. [51] However, to determine whether Tranco would have included the same regionally-popular sites, and thus would have been a reasonable alternative to SimilarWeb for our purposes, we analyzed the percentage of top 50 websites in each EU country (based on SimilarWeb) that appeared in the Tranco list.

We find that Tranco would be an unsuitable replacement for SimilarWeb. Across all 20 countries, collectively, only 18% of the top 50 SimilarWeb websites were included in Tranco top 50. This percentage increased to 24% for Tranco top 100 and 32% for its top 1,000. Figure 2 illustrates this limited overlap between similarweb country specific rankings and Tranco global rankings.

4.2 Identification of First Parties

To identify linked domains owned by the same entity as the site that requests them, we follow a set of simple heuristics. First, we look for AS number match derived from a sequence of DNS resolution (in our local machine) and IP-to-AS lookup from Team Cymru [21]. If two domains resolve to an IP owned by the same AS, we infer that the domains belong to the same company and thus the linked domain is a first party. We similarly label domains as first parties if there is an organization match (using the AS from Team Cymru as an input) in CAIDA's AS2Org database [22]. Domains that are not first parties are then labeled as third parties.

We note that this method will produce a lower bound for third-party domains loaded by the initial target sites and thus present in our data. Since our labeling of third-parties is primarily used for finding known trackers that are missed by existing databases (§ 4.6), we argue that ours is an acceptable method for labeling domains as third parties for the scope of this paper. We do not aim to identify every third party tracker, only to provide a reasonable (conservative) estimate of their prevalence, particularly those hosted in non-adequate destinations, in requests sent from the EU.

4.3 Treatment of Google Domains

BrightData imposes restrictions on queries to Google-owned domains. (The reasons for these restrictions are part of a proprietary agreement between the two companies.) Should these queries be sent by the user, BrightData will automatically route them through a "superproxy," which is not in the ASCP that we intend to query, deeming the results of these queries with little value for our exploration of data localization compliance. Thus, we are forced to exclude Google-owned properties from our set of initial targets.

To identify Google-owned domains, we follow the first party identification heuristics (§ 4.2). Using these techniques, we identify 41 Google-owned sites in the set of top sites from SimilarWeb. From these 41 sites, 28 match one of three patterns: 'google.TLD', 'google.co.TLD' or 'google.com.TLD', where TLD refers to any country's top-level domain, such as '.bg'. We also include youtube.com and news.google.com in this list, as they are well-known Google sites. All but 4 of these, or 24 domains, are present in at least one non-Google-owned site: on average, these sites appear in more than 2,000 DNS requests from other sites–they are embedded in a vast number of non-Google-owned sites in many EU countries. BrightData does allow these queries, where a Google site is loaded by a non-Google site, to be routed through the requested ASCP.

Of the 13 additional sites owned by Google in our set of top sites, 1 is loaded by a non-Google-owned domain. In sum, we are only unable to measure data localization compliance for 16 Google-owned top sites, as the remainder are requested by non-Google sites. These represent less than 1% of our target sites from the previous subsection.

We do acknowledge a limitation of our method (§ 10.1 lists this paper's limitations). By not directly loading the 41 Google-owned sites, we may be missing additional trackers that target EU users. However, this limitation is mitagated since the majority of these sites (26) reference Google search's frontpage for a specific country, a relatively simple website that does not typically embed a large number of non-Google sites. The limitation is further mitigated by the fact that these frontpage Google sites are widely present in non-Google sites, so we are still able to infer their data localization compliance with our method (as noted earlier, BrightData does allow proxies to load Google sites if they are fetched by a different initial target domain).

4.4 Final Sample of Top Sites

To maximize response rates, we attempt to query multiple URLs for each top site. Since a website might be responsive to only 'http' or 'https' requests [45], we attempt to query 'https' first, and if we receive no response, we attempt 'http'. Finally, we note that some sites only respond to queries with 'www.' as a prefix to the TLD+1 domain, for instance, 'www.wikipedia.org' instead of 'wikipedia.org'. In sum, we attempt 4 queries for each top site, with each subsequent query only run if the previous one failed: https://www.website.com, http://www.website.com, https://website.com, http://website.com.

After executing our queries through BrightData in each ASCP, we receive responses from 534 popular sites in 20 EU countries.

4.5 Web Crawls Through BrightData

We "browse" all popular sites in each country and record their response, step (3) of Fig. 1. This data collection occurred in August of 2022. Our aim is to avoid triggering anti-bot/anticrawl measures that (likely most) popular sites implement. To reach this goal, we use a headless instance of Selenium with requests routed through a BrightData proxy: these proxies are set up on real users, and Selenium is a properly configured web browser (not a command-line tool such as curl). In practical terms, we submit HTTP/HTTPS requests to each popular site in each country from all ASCPs identified earlier. The request is sent to the target site through BrightData using a Python proxy handler that is initially set up for each ASCP with authentication information (our user ID and a plaintext passphrase), and the proxy port. A BrightData proxy handler follows this expression–in addition to the previously identified fields, TCC is the two-letter country code of the requested proxy:

```
http://lum-auth-token-country-<TCC>:<passphrase>
@pmgr-customer-<user\_ID>.zproxy.lum-superproxy.io:<port>
```

The output of this stage is a set of DNS requests initiated by the browser, which executes JavaScript and other dynamic content. These requests include the initial target site along with any additional domains loaded by it. These domains are the necessary information for our further analyses. While the remainder of our experiments are based on these DNS requests, step (4) of Fig. 1, we also record the web contents and cookies.

4.6 Labeling Trackers

Tracking sites pose a special concern from a privacy perspective. Thus in our analysis we investigate compliance with data localization by all domains, in general, and by tracking domains, in particular. This is depicted in step (4) of Fig. 1. To label a domain as a tracking site, we use a three-step approach applied to the domains found in the Selenium DNS requests (§ 4.5). First, we intersect the domains with known trackers from the well-established list, EasyList [24]. After manually inspecting the 256 third party domains (§ 4.2) we labeled as non-trackers following this step, we found that the vast majority still appeared to be trackers. Thus, second, we complement EasyList with a well-known list of trackers (with over 1k stars) on GitHub [4]; this process yields an additional 167 trackers. Third, for completeness, we manually inspect the remainder third party non-trackers. We find five additional trackers, four of which are labeled as so because of information in their frontpage or 'about us' section (24media.gr, almatalent.fi, cdn-expressen.se, mailchimp.com), and one from their WHOIS registration (labeled as 'Tech Adverts', amlimg.com).

With our final list of tracking domains, we then manually verify their labeling as third-party trackers. This manual verification is necessary because many sites are incorrectly labeled as first parties with our automated approach because the initial domains are also hosted by a company that itself operates a known tracker (and thus they are delivered to users from the same network, *e.g.*, the same autonomous system). For instance, initial target site *topky.sk* in Slovakia is hosted by Google (server IP: 172.217.2.202), which also operates known tracker *ajax.googleapis.com*. Our automated method thus concludes that the latter is a first-party of the former, when in reality they are likely not.

To verify third-party labels, we thus rely on a manual verification process that leverages three primary sources: Google search of the sites involved and their ownership, *WhoTracks.Me* [40], a well-known database of online trackers, and the Whois database of domain registration information [30]. Given the time-intensive nature of this verification, we restrict this analysis to the known trackers inferred to be in non-adequate countries by our server geolocation method (§ 5).

5 SERVER GEOLOCATION

This section covers our method to geolocate servers.

5.1 Source-Based Measurements

We first obtain a preliminary assessment of where the server is located using RIPE IPmap, step (6) of Fig. 1. This assessment is preliminary since even more accurate geolocation databases can err at the country level. The passive inference provides us with a list of candidate server IPs that might be located in a non-adequate country. In this and further subsections, we aim to identify instances of erroneous inference by IPMap; in particular, we identify those where the server IP is located in the EU or an adequate country but that were inferred by IPMap as being in a non-adequate country. In other words, we look to identify false positives in our identification of potential GDPR violations.

Our initial step to accomplish this goal launching traceroutes towards the servers (hostnames) previously inferred as being located in a non-adequate country, step (5) of Fig. 1. Note that this approach ensures that the DNS resolution is conducted on the same ASCP as the source of BrightData measurements. Then, we identify candidate servers that may be located in non-adequate countries since both the traceroute latency and IPMap support that inferred location. This data collection took place in August of 2022.

Specifically, in this step we look for latency between the EU-based RIPE Atlas probe and the destination server (hostname) that is consistent with latency statistics published by Verizon [60]; this is depicted in step (6) of Fig. 1. Since Verizon does not publish latency data between Latin America and the EU, we rely on wondernetwork.com/pings [61] for these destinations. In both cases, we impose a requirement that the observed latency is at least 90% of the average for that destination. These thresholds vary widely depending on the non-EU and non-adequate destination: Europe (13ms), US (65ms), EMEA region (78ms), Asia-Pacific region (106ms), Latin America (113-166ms depending on the country).

We launch 9,905 traceroutes towards servers in non-adequate countries (as per IPMap).[1] In 9,296 of these cases, we analyze the traceroute latency to the last hop, subtracting the latency from the first hop when possible to avoid increased latency in the last mile, *e.g.*, due to WiFi. In an additional 451 instances we use last hop latency. We exclude 158 traces due to either an unresponsive last hop (28) or latency that is higher to the first hop than the last (130). In 8,488 traceroutes we observe latency that is below our threshold for that destination, and we exclude these from further investigation. We are left with 1,259 traceroutes that suggest that a server is located in a non-adequate country; recall that the European Commission has designated a number of non-EU member countries as being "adequate" for the purposes of the GDPR's data localization requirement. [13]

[1]We launched traceroutes also towards servers in EU and adequate countries; we do not analyze these traceroutes here, as they are unlikely to contribute information on destinations in non-adequate countries. We will release these traceroutes with the rest of the data and code in this study.

5.2 Destination-Based Measurements

To further confirm that responding servers are located in non-adequate countries, we collect additional evidence from RIPE Atlas probes located in those same countries. This data collection took place in November/December of 2022. We then use speed of light (subsequently denoted by c) constraints to discard likely erroneous geolocation inferences by IPMap. Our goal is to remove as many false positives as possible from our experiments using empirical network data. However, there may still be false positives in our results; we describe this limitation of our work in § 10.2.

We launch traceroutes from RIPE Atlas probes located in the same non-adequate country where the server was inferred to be located by IPMap, step (5) of Fig. 1. In this case, the destination is the IP address of the server rather than a hostname, as the DNS resolution was already done from the same network in § 5.1. We analyze the latency to the last hop, subtracting the latency from the first hop as before. We launch 598 measurements; we only measure each destination IP once from each AS-country pair–with the AS being the same as that from the source-based measurements and the country being that inferred by IPMap for that IP–regardless of how many times the destination IP appears in the source-based measurements.

We exclude 19 measurements, step (6) of Fig. 1, due to unresponsiveness of either the last hop or the RIPE Atlas probe, and 57 due to insufficient granularity in the RIPE IPmap inference (or the probe's location) to compute geodesic distance. Of the remaining 522 measurements, in 385 cases we rely on the difference in latency between the last and first hops, and use the last hop latency in all others. Of these, 130 exhibit higher latency to the first hop than the last, a contradiction that we may be caused by additional (home) router delays due to the generation of an ICMP response, compared to forwarding an incoming ICMP message from another device. Unlike in § 5.1, we keep these measurements here as by now we have at least three pieces of evidence that the server is in a non-adequate country, decreasing the likelihood that the server is located in the EU (recall that our goal is to remove as many false positives as possible, as that would erroneously indicate a potential GDPR violation). In 7 additional cases, the latency to the first hop is not available (router did not respond to ICMP request).

We then infer whether this latency is consistent with the geodesic distance between the RIPE Atlas probe and the destination IP as inferred by RIPE IPmap. To account for the Internet's non-geodesic routing due to physical constraints, such as the speed of light in fiber being $2c/3$ [33], or infrastructure delays, such as queue buildups on routers, our upper bound for observed speed is $4c/9$ [32] or approx. $133km/ms$; this is a more conservative threshold than the frequently used $2c/3$. If the speed inferred from the traceroute round-trip travel time and the geodesic distance between the endpoints is higher than $4c/9$, we discard the measurement, which happens in 89 instances. We then have 433 measurements remain that target servers still inferred to be in non-adequate countries.

5.3 Reverse DNS Lookups

As a final piece of evidence in our server geolocation methods, we inspect reverse DNS (rDNS) records of each traceroute's last hop (reported by RIPE Atlas), step (7) of Fig. 1. Hostnames obtained from rDNS are often, but not always, useful in geolocating IP infrastructure [37], which is why this is the last step in our analysis.

Of the 433 measurements from the last subsection, 255 include hostnames that confirm the server's country inferred in previous steps (206 of these refer to servers in the US). For instance, hostname **unn**-138-199-8-197.datapacket.com most likely refers to IP infrastructure near Ranong Airport (IATA code: UNN) in Thailand, which is the same country as inferred by IPMap for the corresponding server's IP address. Given the diversity in operational practices to assign hostnames to IP infrastructure, it is not trivial to automatically infer geographic hints to determine where the referenced infrastructure is located; our re-implementation of recent work seemed to miss some geographic hints in hostnames [37], which is why we manually inspect all the hostnames in this step - an effort that is supported by the data's manageable scale.

The rDNS records for a further 13 traceroutes suggest that the server is located in a different non-adequate country than that inferred by RIPE IPMap. In these cases, we reassign the IP to the non-adequate country inferred from rDNS (which tends to be more accurate than latency-based inferences). Furthermore, the hostnames for 37 measurements suggest that the servers are located in either the EU itself, or an adequate third-country. Nearly all of these (31) refer to AWS infrastructure that seems to be located in Canada but were erroneously inferred by IPMap to be in the US, *e.g.*, ec2-99-79-143-255.**ca-central-1**.compute.amazonaws.com. We exclude these 37 IPs from further processing, as these servers are unlikely to be located in a non-adequate country (recall that Canada is an adequate country [13]).

Finally, 45 measurements do not return a hostname with the rDNS lookup, and another 83 do not seem to encode geographic locations. We keep these servers' location inference unchanged from previous steps.

5.4 Final Sample of Non-Adequate Servers

We are left with 396 measurements to 247 server IP addresses where all available evidence suggests that the server responding to EU requests is located in a non-adequate country. These potentially non-adequate servers are present in our data in 1,233 instances, as the same server may be contacted by multiple initial sites in each ASCP. Table 1 shows all methods we used to investigate the geolocation of servers.

6 ANALYSIS

We now present our analysis of EU collected data. We concentrate our investigation along several directions. First, we investigate how frequently EU requests are served from non-adequate destinations in each source country. We further analyze the prevalence of known third-party trackers that are located in non-adequate destinations by source country. While both of these scenarios constitute potential GDPR violations, the latter is a stronger case since known trackers are more likely to be capturing personal information and by definition are sending such data to third parties.

Table 1: Sever Geolocation. *No hostname/no geohint. **RIPE IPMap inferences.

Method	Probes	Unresponsive	Adequate	Non-adequate
Source traceroutes	9905**	158	8488	1259 (598 IPs)
Destination traceroutes	598	76	89	433
Reverse DNS	433	45/83*	37	396

Second, we analyze the most prevalent trackers, and the types of websites that load them, in each source country. Third, we investigate whether there are regional differences across the EU in compliance with data localization. Finally, we study the cookies loaded by sites that contact non-adequate servers.

6.1 Servers in Non-Adequate Countries

We now turn to the question of prevalence of servers, generally, and tracking servers, specifically, in non-adequate countries by source EU country. Tab. 2 shows a summary of three key metrics in this regard. First, the percentage of traceroutes sent from each EU country that reached a server in a non-adequate country. Second, the percentage of unique server IPs that are hosted in a non-adequate country, and third, the same metric for unique tracking servers.

At a high level, we find that data localization is complied with in the vast majority of cases, as evidenced by the low percentages in all columns of Tab. 2; we present the last two columns as a geographic heatmap in in Fig. 3. (In § 6.6, we investigate the apparent geographic trends in Fig. 3.) However, since we collected data from the most popular sites in each country, the exceptions still represent potentially very large traffic volumes from EU countries to non-adequate countries. For instance, 4.6% of unique IPs that serve users in Poland are located in a non-adequate country, and the same figure is 4.2% for Greece. Moreover, in a majority of the 20 EU countries we studied, more than 1% of third-party trackers are hosted in non-adequate destinations. (We further investigate these trackers in § 6.3.)

6.2 Destination Countries

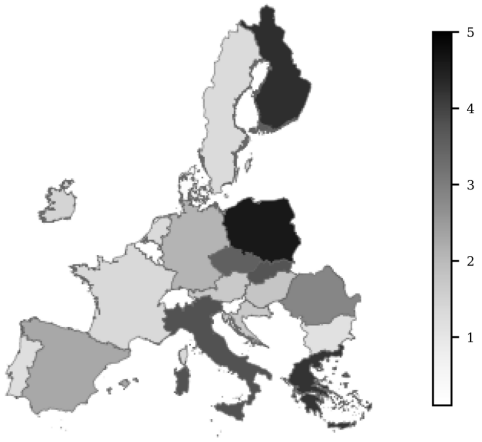
The non-adequate destination countries in our sample span several continents. In Tab. 3 we show those most commonly occuring among these countries. The US, Turkey and Russia account for approximately 90% of traceroutes crossing from the EU into a non-adequate destination. While the US is a relatively common destination for most EU source countries, Russia and Turkey are much more prevalent in two nearby source countries: Finland and Romania, respectively.

6.3 Trackers

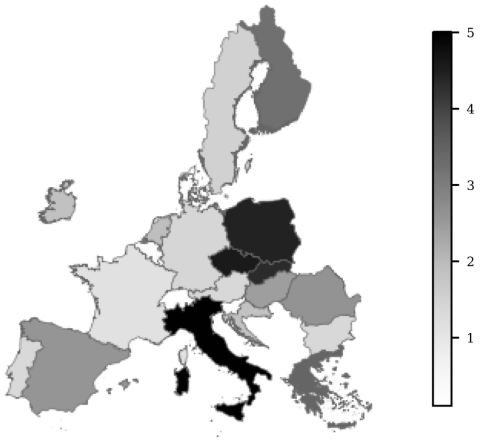
We now turn our attention to known third-party trackers. As stated earlier, these domains are of particular concern from a privacy standpoint because of their collection of sensitive information from users,

Table 2: Percentage of traceroutes, servers and trackers reaching non-adequate destination countries, by source EU country, sorted decreasingly on row average (not shown).

Source Country	Traceroutes Unique	IPs	Unique Tracker IPs
RO	6.6	2.8	1.4
FI	4.3	4.3	1.1
IT	0.7	3.8	4.0
GR	1.4	4.2	2.9
SK	0.8	3.7	3.7
PL	0.8	4.6	2.4
CZ	0.3	3.5	0.8
ES	0.4	2.3	1.6
HR	0.9	1.7	1.6
HU	0.4	1.8	2.0
AT	0.6	1.7	1.2
DE	0.6	2.1	0.5
PT	0.4	1.2	1.3
NL	0.2	1.3	1.4
IE	0.9	1.4	0.5
BG	0.2	1.1	1.4
FR	0.4	1.3	0.4
SE	0.3	1.3	0.4
DK	0.1	0.2	0.0
BE	0.1	0.1	0.0



Percentage of Unique Server IPs in Non-Adequate Destinations
(a)



Percentage of Unique Tracker Server IPs in Non-Adequate Destinations
(b)

Figure 3: Percentage of unique tracker IPs in non-adequate countries.

especially in cases where it is transferred to non-adequate countries. We investigate three aspects of the trackers in non-adequate countries observed in our data: the source-destination pairs of countries, the most commonly occurring trackers in each source EU country, and the most common types of websites that host the trackers.

6.3.1 Source-Destination Country Pairs. In Fig. 4 we show the pairs of source EU countries and non-adequate destination country in our data; each flow is weighted by the number of traceroutes seen between each pair of countries. As observed before for traceroutes towards all destinations (not just trackers), the US is also the most common destination among trackers in non-adequate countries. The US is the most common destination from all but five EU countries in our sample: Netherlands, Ireland, Czech Republic, Romania and Finland.

Four destination countries are also notable in Fig. 4. India is observed as a destination from 13 EU countries; in over 80% of cases (among 49 traceroutes), a Google (third-party) tracker is responsible for this. Russia, Turkey and Hong Kong are frequent destinations from three EU countries: Finland, Romania and Greece, respectively. Thus, the trend observed earlier for all traceroutes from the former two EU sources holds also for known trackers. A majority (26/43) of the traceroutes between Romania and Turkey are caused by a single popular website: *fandom.com*, a site focused on arts and entertainment [55]. In the case of Finland and Russia, exactly half (30/60) of the traceroutes are caused by popular website *vk.com*, a social media network [55] that is popular among Russian speakers. (Russian speakers are a minority group in Finland, which borders Russia). Finally, 17 of 19 traceroutes from Greece to Hong Kong are caused by a single third-party tracker: Facebook, whoose trackers are loaded by 10 popular websites. *makeleio.gr*, a news and media publisher [55], is the most commonly occurring popular website, responsible for 4/17 of these traceroutes.

Table 3: Number of traceroutes reaching top 10 non-adequate Countries (together they account for 97.7% of these traceroutes).

Source → Destination ↓	RO	FI	GR	HR	IE	PL	SK	IT	AT	DE	PT	FR	ES	HU	CZ	SE	NL	BG	BE	DK
US	3	3	16	29	34	85	46	47	43	41	5	12	58	13	17	16	16	7		
TR	308	2		2	11											11	2	1		1
RU		147	2						1				1			7		2		2
MX						35		2				1	1		1					
IN	1	1	4	1	4	4	2	5	4	1	2		3		1		1	1		
SG			6			1			2	1		1	1		2		7			
HK	1		7	1	1			1		5	1		1	1	1					
BR	2						1	5				1	1		2		1	2	2	
AE								3	1			1	1		4			1		
AU						1		1		2			1		1					

6.4 Common Paths

We now investigate common paths that take a request from an EU country to a non-adequate destination. We define a path as a combination of source country, initial website, tracker, and destination country. Among these paths, we show those that occur at least 5 times in our traceroute set in Fig. 5. Two of these paths, which originate in Romania and Finland through *fandom.com* and *vk.com*, respectively, were mentioned earlier. (*userapi.com* might be "affiliated" with *vk.com*, but we could not confirm the former's ownership by the latter, thus we classify this site as a third-party.)

At a high level, we observe a variety of paths that are not particularly concentrated among any one website or tracker. Notably, only four of these trackers seem to be operated by major US companies: *dailymotion.com*, *gvt1.com* and *gstatic.com* (Google), and *clarity.ms* (Microsoft). As before, there are a wide variety of paths that reach the US through combinations of popular websites and trackers.

6.5 Initial Website Categories

We now turn to the question of which types of websites are responsible for loading these third-party trackers in non-adequate countries. Anectodally, news sites seem prevalent among these; for instance, in Fig. 5, all the initial websites in Spain are news websites. We now systematically investigate whether that is broadly the case. In Fig. 6 we present the *SimilarWeb* categories [55] associated with the 239 websites that load at least one of these trackers. A slim majority of these sites, 120, belong to the News & Media Publishers category (including the four aforementioned sites in Spain). This category makes intuitive sense as a frequent fetcher of trackers due to the news industry's increasing reliance on digital advertising revenue and thus on web tracking. Nevertheless, the high prevalence of this category in the set of sites loading trackers in non-adequate countries is still notable. For instance, three other categories include at least 20 websites: Arts & Entertainment, Computers Electronics and Technology, and Ecommerce & Shopping. However, none of them come close to the prevalence of News & Media Publishers.

6.6 Regional Variation

In previous subsections, especially in Fig. 3, we anecdotally observed that the rates of servers located in non-adequate countries seemed higher in Southern and Eastern Europe. To evaluate whether this is true, systematically, we use the United Nations defition [23] for four regions of Europe: Northern, Southern, Eastern and Western. We evaluate both the rate of server IPs, overall, and tracker IPs, specifically.

The findings are shown in Fig. 7. The rates of presence in non-adequate countries is higher in Southern and Eastern Europe, compared with Northern and Western Europe, for both server IPs and tracker IPs. We use an ANOVA test to determine whether these regional differences are statistically significant. For server IPs, they are not ($p = 0.19$). For tracker IPs, however, the difference is significant ($p = 0.01$). This latter category presents bigger privacy risks. This disparity can lead to higher risks of privacy harms for users in Eastern and Southern Europe.

6.7 Cookies

In this subsection, we present an analysis of the cookies that were loaded by initial sites that contacted at least one server in a non-adequate country. We find substantial evidence that servers in non-adequate countries are engaging in user tracking activities.

Of the 1,233 non-adequate instances observed in § 5.4 (recall that an instance is an initial site loaded from an ASCP), we find that 824 retrieve and store non-empty cookies. Of these, 236 websites load cookies with unique identifiers, for a total of 9,885 cookies. Unique identifiers pose a potential privacy harm, as they can be used to track users beyond the website they are currently browsing. These cookies contain 1,153 unique identifiers. The unique identifiers [42] in a cookie's name or value can provide valuable context about the organization that issued or uses the cookie. We also used Cookiedatabase [18] for identifying organizations based on the unique identifier. We found 494 cookies that contain an identifer `_ga`, which indicates they were set by Google Analytics. Similarly, 457 cookies have `_gid` and 254 have `__gfp_64b`, which

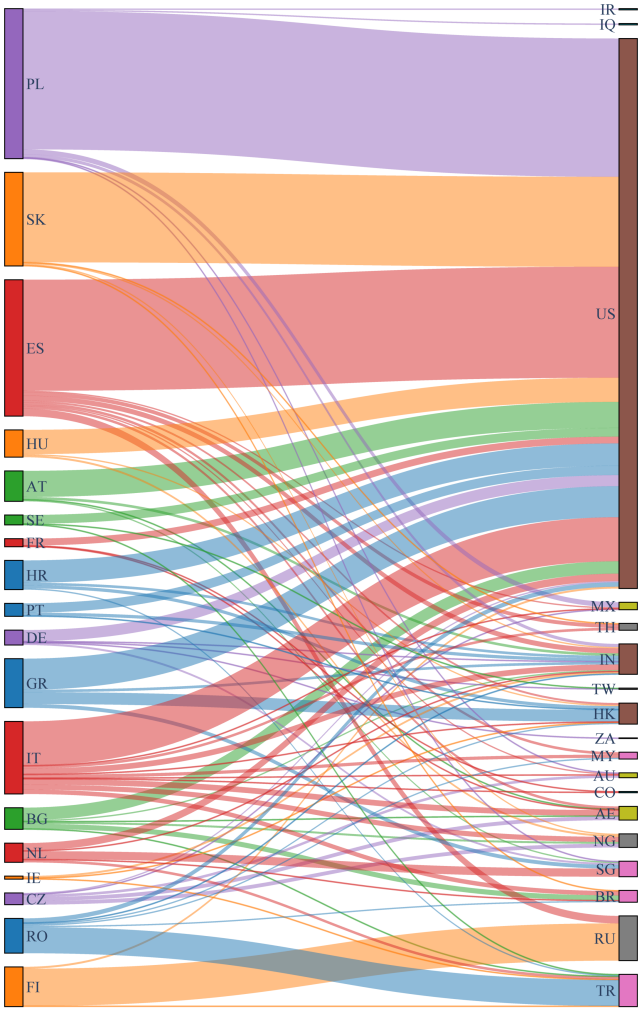


Figure 4: Flowchart showing, among known third-party trackers located in non-adequate countries, the prevalence of traceroutes connecting each source-destination pair of countries.

indicates they are used by Google services for analytics and ad personalization (DoubleClick), respectively.

Other popular cookie identifier were, `_fbp` (230 cookies), set by Facebook for marketing. Additionally, we observed various identifiers that are related to consent management such as `\_pbjs\_userid\_consent\_data` and `OptanonConsent`. Consent management ensures compliance with privacy regulations by enabling users to control their data collection preferences. Figure 8(a) illustrates the distribution of the most frequently observed cookies.

Our analysis also identified the most commonly occurring cookies across websites. The most frequent cookie was `_ga` (Google Analytics), found on 146 websites, followed by `_gid` (Google Analytics) on 135 websites, `__gfp_64b` (Doubleclick) on 84 and `_fbp` (Facebook) on 63 websites. Figures 8(b) illustrates and highlights the prevalence of tracking and analytics cookies across websites.

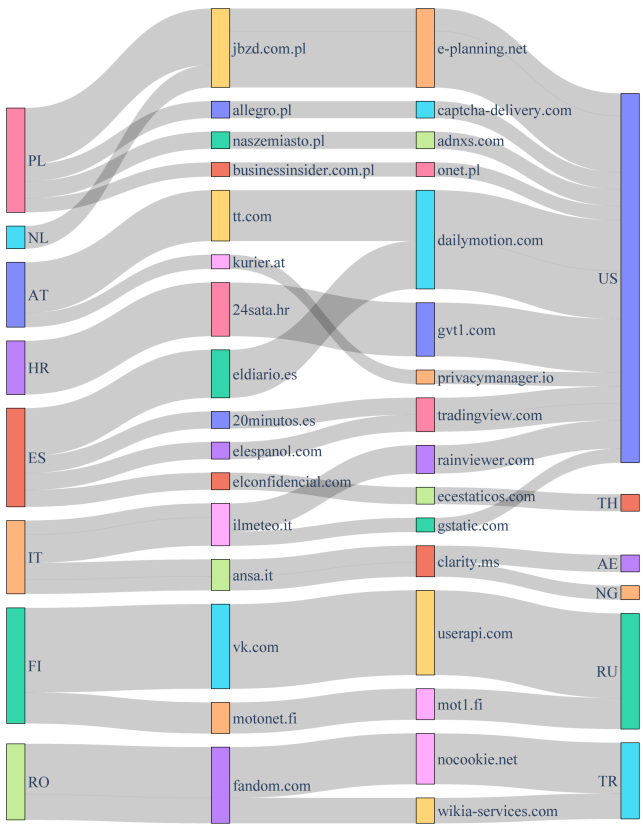


Figure 5: Combinations of source country, initial website, third-party tracker, and destination country that are observed five times or more in our data.

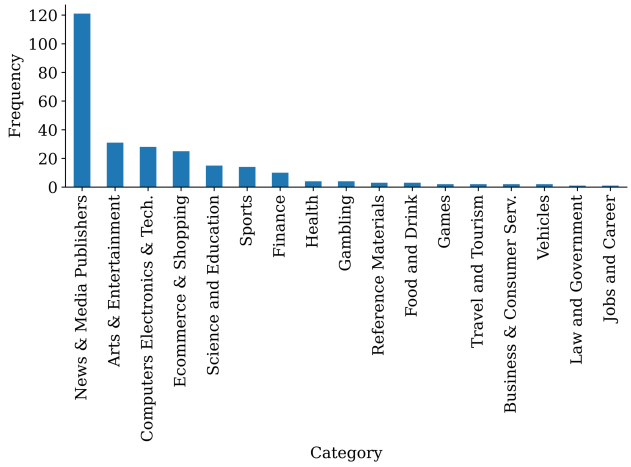


Figure 6: Categories for websites that load observed trackers in non-adequate destinations.

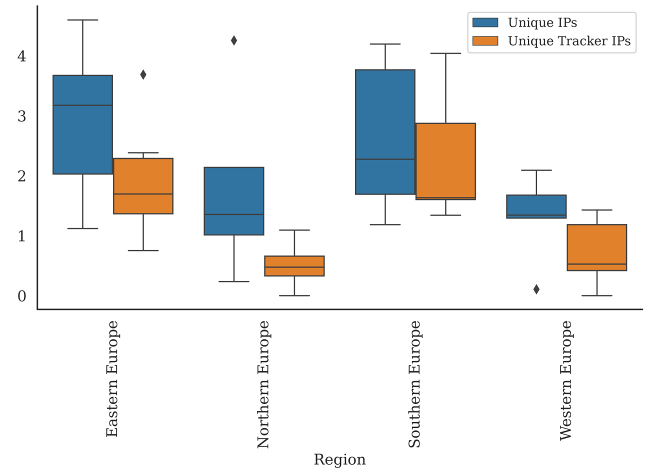


Figure 7: Boxplots showing the percentage of non-adequate server IPs and tracker IPs for the countries present in each EU region.

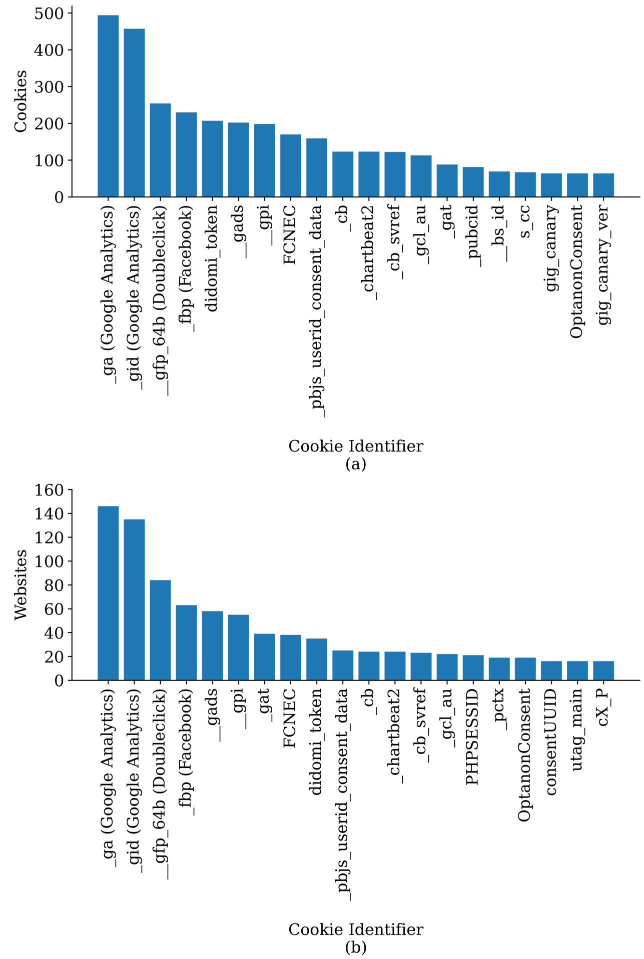


Figure 8: (a) Cookie identifier associated with the number of cookies. (b) Cookie identifier and number of website they are embedded.

7 DISCUSSION

In this section we discuss our findings on data localization compliance in the EU.

7.1 Causes for Servers/Trackers in Non-Adequate Destinations

Our method does not reveal the causes for web servers or trackers to be located in non-adequate countries, which is potentially unlawful in the EU. We speculate about some potential causes here. First, the content provider is intent in serving EU users from the EU (or an adequate country), but in the presence of temporary server or router outages, it directs the user request to a backup location elsewhere. Second, the content owners may not know that the user is located in the EU, as Internet protocols were not designed with physical nor jurisdictional constraints. For instance, *vk.com* might mistakenly infer that a user is located within Russia, thus directing them to their servers there. Third, in some instances, performance considerations may trump legal compliance. For instance, users in Romania might experience much higher latency if their content is served from a server in, say, Belgium, rather than Turkey, which is much closer.

7.2 US as Destination

The US is the most frequent non-adequate destination in our data, observed from all but two EU countries in the sample. The US is, of course, a key provider of cloud services, and the home base of most of the world's largest content providers. Partly given the amount of economic activity spurred by data transfers between the US and the EU, the executive branches of both jurisdictions are intent on generating a framework that broadly authorizes such transfers [14, 29]. Such a framework is not guaranteed to stand up in the EU courts as previous attempts at authorizing transfers from the EU to the US have failed in the EU's legal system [36]. Whether the latest EU-US framework will pass muster under future judicial review remains an open question according to legal experts [11].

7.3 Regulatory Considerations

The wide variety of trackers observed, loaded from an also varied set of popular sites in each country, may pose challenges of scale for case-by-case regimes such as the Standard Contractual Clauses (SCCs) [15], particularly when auditing compliance. These have been applied to US companies [27] and are also in consideration [15] for data transfers to countries in Southeast Asia (ASEAN). We observed multiple ASEAN member countries [2] as destinations in our data: Thailand, Singapore and Malaysia. While they are relatively uncommon destinations in our sample, EU-ASEAN agreements [15] might change that in the future.

7.4 News Sites

The concentration we observed, where trackers in non-adequate countries are primarily loaded by news sites, provides a potential

opportunity when designing data localization auditing frameworks in other jurisdictions outside the EU: the websites in this category should be evaluated first. This prioritization might be particularly useful in pilot evaluations of compliance with newly introduced data localization requirements, or in environments where data collection is constrained for instance by available bandwidth or where electric supply is less reliable.

7.5 Regional Differences

Perhaps the most surprising result in our study is the differing rates of compliance across EU regions. While certainly there are cultural and economic differences between them, in theory it is not more challenging for a content provider to deliver content from within Europe: there is ample available infrastructure within the continent. However the disparities in compliance with data localization principles for trackers were significant. This uneven compliance with regulation intended to protect user privacy poses questions in equity and fairness: the wealthier regions of the EU (Northern and Western) might be able to provide more uniform protections to their users than the less affluent regions in the south and east. Extrapolating this to other jurisdictions who may be considering, or have recently implemented data localization requirements, these equity and fairness considerations are worth examining empirically. Empirical audits as the one we have presented in this study can reveal these issues with the greatest clarity.

7.6 Russia and Finland

The case of *vk.com*, as accessed in Finland, presents a particular challenge for data privacy. The company's CEO has been sanctioned by the EU in connection with the Ukraine conflict [41]; the company is also state-owned (by Russia) [5]. However, since the services provided by VK are popular among Russian speakers, which are a substantial portion of the population of many Eastern EU countries and Finland, an outright ban of the service would negatively impact these groups. We do find that *vk.com* loads third party trackers in Russia, creating a potential geopolitical challenge in addition to the usual potential privacy harms.

8 SERVER GEOLOCATION VALIDATION

To validate our server geolocation method, we build an experiment based on two sets of server IPs with known location. Each of these sets is based either in a non-adequate country, the US, [2] or an adequate country, an EU member state. To conduct this validation experiment, we followed the four steps of the methodology defined in § 5 to perform server IP geolocation: RIPE IPmap geolocation, source-based constraints, destination-based constraints, and reverse DNS (rDNS).

The results of this experiment are shown in Tab. 4. In the following subsections, we describe this validation experiment. We also infer rates of true/false positives and precision of our method. Note that, as discussed in the limitations section (§ 10.2), the rate of false positives might be different in our main results.

[2] At the time of our main data collection. The validation experiment was conducted in Feb. 2025, when the US had already received an adequacy decision. [34]

8.1 Server Testbed in the US

We have access to a US testbed, CloudLab [12], and thus to ground truth mapping between server IP and country of operation. We conducted experiments using 200 such servers from this testbed that are distributed across two different locations (L1 and L2), both in the US. These servers are the destination IPs for this experiment. Further, we selected five source countries from the European Union (EU). We included the largest EU country, France (by land area), and four additional countries chosen randomly to represent different regions of the EU: Poland from Eastern Europe, Germany from Western Europe, Ireland from Northern Europe, and Spain from Southern Europe. For each source country, we randomly selected 40 server IPs as destinations across the two testbed locations (L1 and L2). This results in a total of 200 destination IP-source country pairs (DSCPs).

The upper half of Tab. 4 shows the results of this experiment. In the rest of this paragraph, we describe these findings in more detail. We found that RIPE IPmap did not provide country-level geolocation for three server IPs. In the source-based traceroutes, five server IPs failed to respond, and 20 did not meet the 90% latency threshold described in § 5.1. We further excluded two server IPs using destination-based constraints. Finally, when we analyzed rDNS data, three of these servers had clues indicating they were not located in the city identified by RIPE IPmap, but they were still located in the US. As our method is concerned with country-level geolocation, we keep the server geolocation unchnaged from previous steps. In summary, our final dataset correctly geolocated 170 DSCPs as being in the US, and discarded 30 DSCPs due to the constraints described above.

8.2 AWS Servers in the EU

To validate our method for adequate server IPs, we conducted an experiment using 1,000 IPs in the EU advertised by Amazon Web Services [54]. We selected these IPs by intersecting the AWS-published IP ranges with the ISI IP Hitlist, [1] which estimates the likelihood that a server will respond to network measurements. The IPs we select have a score of 99 on the Hitlist. We take this step in an attempt to maximize response rates to our measurements since, unlike on the US testbed, we have no direct knowledge of whether the AWS IPs are currently both routed and in use. This is also why we increase the number of server IPs in the EU relative to the US experiments in the previous subsection.

Similarly to our treatment of the US testbed, we divide the server IPs into five groups corresponding to the same five EU source countries. The lower half of Tab. 4 shows the results of this experiment.

From these 1,000 DSCPs, we find that RIPE IPMap does not assign a country in 36 cases, and erroneously geolocates 270 DSCPs outside of the EU in non-adequate countries: one in Pakistan, the rest in the US. After conducting both our source- and destination-based measurements, and applying speed of light constraints, these 270 DSCPs are all discarded. At this stage, 27 DSCPs are correctly labeled as being in an (adequate) EU country. rDNS confirms this assertion in all 27 cases.

Table 4: DSCP Geolocation for server validation experiment. ST/DT: Source/Destination Traceroutes. TB: Server Testbed. *No hostname/no geohint.

Method	DSCPs	Unresponsive	Adequate	Non-adequate
ST: US TB	197	5	0	172
DT: US TB	172	0	0	170
rDNS: US TB	170	99/0*	0	170
ST: AWS	964	910	50	0
DT: AWS	50	21	27	0
rDNS: AWS	27	0/0*	27	0

8.3 True/False Positives and Negatives, Precision

In this subsection, we describe the rates of true/false positives and negatives, along with precision, that result from our validation experiments.

False positives and true negatives. The outcome of the AWS experiment is either a false positive, where a server in an adequate country is incorrectly inferred as being in a non-adequate country; or a true negative, where a server in an adequate country is correctly labeled as being so.

Given the results of the AWS validation experiment, we infer *no false positives*. This is because all servers initially labeled as being outside the EU by RIPE IPMap were correctly discarded—assigned a negative label—by subsequent steps in our method. Note that our source- and destination-based filters aggressively discard DSCPs because they are either unresponsive or the latency fails our speed of light constraints. Thus, after all filters are applied, the *true negative* rate is 100%.

In aggregate, these results suggest that the coverage of our method is limited, because it excludes servers that may actually be located in a non-adequate country. Simultaneously, our method reduces the likelihood of false positives. Thus, in this experiment, our method is working as intended.

True positives and false negatives. Recall that the US was a non-adequate destination at the time of our main data collection, so all servers in the US testbed have a non-adequate country as their ground truth location; thus, all DSCPs in the US testbed validation experiment are either correctly allocated to the US, a true positive, or discarded, a false negative. Our method correctly identified 170 DSCPs as being in the US, of 200 DSCPs known to be in the US. Therefore, the *true positive* rate of our method in this experiment is 85%, and the *false negative* rate is 15%.

Precision. From the aforementioned results, in particular given the absence of false positives, the inferred *precision* of our method in the validation experiments is 1.0.

9 PROXY LOCATION VALIDATION

We conduct an experiment to investigate whether BrightData's claims about requests being routed through an AS-Country Pair are accurate. To this end, we set up a web server at Northeastern University and send HTTP requests through BrightData from each ASCP. All of the requests this server received were IPv4, and we take steps to preserve the privacy of BrightData users (who host the proxies in their own devices) by recording only the /24 subnet from which we received the request to our university server. We then compare the country and AS claimed by BrightData with those identified by geolocation database Maxmind [38]. We fetch the AS and country for every IP in the /24 prefix through Maxmind.

We find that BrightData seems to be almost always routing requests through the ASCP they claim. Of the 2,319 valid requests received by this server from BrightData, all but five are accurate. Thus, 2,314 requests have an IP that is part of a /24 prefix entirely present in the same ASCP according to Maxmind. The five exceptions include two where the country does not match (but the AS does), two where the AS does not match (but the country does), and one where neither AS nor country are a match. Therefore, we conclude that BrightData is an appropriate proxy to use for the purposes of routing requests through a specific AS in a given country.

We acknowledge that geolocation databases are prone to errors. However, since we are working at the country level granularity, these errors are less common [48, 52]. Of course, it is possible that both BrightData and Maxmind are often both incorrect and in agreement about the ASCP where a user is located, but we argue that this is a remote possibility.

10 LIMITATIONS

In this section, we describe the main limitations of our approach.

10.1 Domain Exclusions

As described in § 4, our framework excludes both Google-owned and adult websites as initial domains. This exclusion is caused by BrightData rules. Thus, our method is not able to study these two groups of domains, particularly as initial sites. For Google domains, this limitation does not apply when they are loaded as third-parties by other websites that are not Google owned.

10.2 Potential for False Positives

Since we do not have access to ground truth on physical server location, the rate of false positives in our results is not known. For instance, there could be interactions between the server response rate to traceroutes and their geographic location, which would impact our findings. While we conducted a validation experiment using servers with known location (§ 8), it is still possible that the servers loaded by popular EU sites, *i.e.*, the servers we study in our main results section, respond differently to our measurements than the servers in the validation experiments. Furthermore, the EU servers in the validation experiment may treat traceroutes differently from other traffic, potentially impacting our latency-based geolocation techniques. Finally, the scale of our validation experiment is smaller than the experimental setup in our main results, which may impact the former's generalizability. Therefore, the false positive rates in the validation experiment might differ from those in our main results. Due to all the aforementioned factors, the lack of information regarding false positive rates and precision in our main results is a limitation of this study.

10.3 Alternative CDN Nodes

A key limitation of our study is that we do not collect alternative server locations for each destination, for instance available CDN nodes [56] from each ASCP. This additional CDN-node data would potentially reveal whether the content providers are making their best effort to comply with GDPR's data localization policies, *i.e.*, they are picking compliant servers even when they offer worse performance. Alternatively, the content providers could instead primarily be selecting these servers based on performance attributes such as latency. Thus, our study is not able to distinguish between server placement that is based on performance constraints or GDPR data localization compliance.

10.4 Inference of GDPR Violations

Our study is not able to make final determinations on GDPR violations. This is because legal exemptions may exist to mitigate the instances where we have identified servers in non-adequate countries. Further, a ruling on any potential GDPR violation would require additional context on the specific data that was transferred and any legal contracts in place between various entities, including privacy regulators [15]. Ultimately, as such a determination of a GDPR violation would likely take place in a judicial context and be subject to considerations beyond the physical location of a server.

10.5 Framework Applicability in Other Regions

Our framework has limited applicability beyond the EU due to measurement infrastructure density. Previous work has found that RIPE Atlas has more density of deployment in Europe compared to other regions [47, 53]. While geographic bias in the BrightData platform has not been as rigurously quantified, even a cursory look at their top proxy locations reveals potential bias, with the US and India having a similar number of IPs available even though the latter is considerably larger in terms of Internet users. [9]

11 RELATED WORK

Iordanou et al. [31] studies the cross-border web tracking that targets users in the EU, specifically the geographic scope of these data flows. There are some similarities between this study and our work. As in our study, the previous work leverages large-scale Internet measurements, partially launched from real end-user devices, to study the location of servers responding to requests in the EU. Both the related work and our study conclude that most traffic stays within the EU (or, in our case, in the EU and adequate third countries); and both also rely on RIPE IPmap for server geolocation, though our framework also collects both source- and destination-based measurements, and rDNS data, to validate each individual server inferred to be in a non-adequate country.

However, this previous work [31] differs from ours in three key ways. First, their focus is entirely on tracking servers, as they are interested in tracking flows, whereas we investigate both general-purpose servers as well as tracking servers, as we are instead interested in auditing data localization. Second, the related work's focus is on maximizing coverage of potential *tracking flows*, whereas our study is focused on obtaining a representative sample of *source networks* in the EU. Third, Iordanou et al. launch measurements from a browser extension deployed to 183 EU users over a longer period of several months, and intersect the tracking servers there observed in ISP NetFlow data from four large networks in three countries, whereas we rely on a point-in-time collection from a proxy deployed to more than 1,000 networks in 20 countries.

Previous work has also tackled the issue of identifying cross-border data transfers. Guamán et al. [28] study the geographic spread of data flows originating from Android apps and evaluates whether these flows comply with both the developer's privacy policies as well as the GDPR. Razaghpanah et al. [49] analyze tracking by mobile apps and whether these flows comply with EU regulations, including an analysis of non-EU server locations. Similarly, Nan et al. [43] apply static analysis techniques to IoT companion apps (the smartphone apps needed to operate most IoT devices) and reveal data exposure caused by these devices. Finally, Urban et al. [59] study third-party services in the Web to generate a tree of dependencies that represents which additional services are loaded by each initial third party contacted by a website. While not a central focus of the latter two studies, the authors do investigate the country where third-party servers that receive IoT/Web data (respectively) are located. We note that all the studies cited in this paragraph rely on a geolocation database (prone to inaccuracies [48, 52]) to pin the location of servers, a method which often overstates the prevalence of servers located in the US [31], without conducting additional verification using Internet measurements, as we have done in our work. Thus, we argue that our framework is more applicable for auditing compliance with data localization requirements.

12 CONCLUSION AND FUTURE WORK

A key component of the GDPR is the data localization regulation. However, to date there was not a method to audit compliance with this requirement at continental scale. Our method fills this gap and provides a framework for empirically auditing compliance with data localization principles and laws, and specifically investigate such compliance for known trackers, which pose a greater privacy risk. To accomplish this, we collect both browser-based data, the websites and domains loaded while browsing popular sites, and network-based data, the servers that respond to such requests. We find that data localization requirements are broadly complied with in the EU, though there are meaningful exceptions, which are more prevalent some regions.

In the future, we plan on investigating the causes of high compliance rates, disambiguating between legal compliance and performance prioritization by web content companies. We further plan on applying this framework to regions beyond the EU. In these regions, there are regulations that differ from the GDPR on data localization to varying extents, providing opportunities for comparative analyses of policy effectiveness with regards to data localization.

13 ACKNOWLEDGEMENTS

We thank the anonymous reviewers and the revision editor for their valuable feedback. We are grateful to BrightData and RIPE Atlas for allowing us to collect data using their platforms. This research was funded in part by the US National Science Foundation (NSF), Grants No. CNS 1955227 and CNS 2402963. Author Gamero-Garrido was supported in part by Northeastern University's Future Faculty Fellowship and the Ford Foundation Postdoctoral Fellowship.

REFERENCES

[1] USC Information Sciences Institute ANT Project. 2024. IP Hitlist Dataset. https://ant.isi.edu/datasets/ip_hitlists/format.html Accessed: February 28, 2025.

[2] ASEAN. 2024. ASEAN Member States. https://asean.org/member-states/.

[3] RIPE Atlas. 2021. Probe Archive. https://ftp.ripe.net/ripe/atlas/probes/archive/2021/12/20211101.json.bz2. (Data for November 2021.).

[4] badmojr. 2022. 1Hosts. https://badmojr.gitlab.io/1hosts/. (Fetched in August 2022.).

[5] The Bell. 2024. Russia takes direct control of top social media networks. https://en.thebell.io/russia-takes-direct-control-of-top-social-media-networks/.

[6] European Data Protection Board. 2020. Frequently Asked Questions on the judgment of the Court of Justice of the European Union in Case C-311/18 - Data Protection Commissioner v Facebook Ireland Ltd and Maximillian Schrems. https://www.edpb.europa.eu/sites/default/files/files/file1/20200724_edpb_faqoncjeuc31118_en.pdf.

[7] European Data Protection Board. 2025. International data transfers. https://www.edpb.europa.eu/sme-data-protection-guide/international-data-transfers_en.

[8] BrightData. 2023. BrightData. https://brightdata.com/.

[9] BrightData. 2023. Top Proxy Locations. https://brightdata.com/locations.

[10] Duc Bui, Brian Tang, and Kang G Shin. 2022. Do opt-outs really opt me out?. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 425–439.

[11] CEPS. 2024. In the EU-US data transfer and privacy quarrel, the end is not in sight. https://www.ceps.eu/the-eu-us-data-transfers-and-privacy-quarrel-the-end-is-not-in-sight/.

[12] CloudLab. 2025. CloudLab. https://www.cloudlab.us/

[13] European Commission. 2022. Adequacy decisions. https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en. (Accessed on 02/20/2023).

[14] European Commission. 2022. EU-U.S. Data Privacy Framework, draft adequacy decision. https://ec.europa.eu/commission/presscorner/detail/en/qanda_22_7632. (Accessed on 02/20/2023).

[15] European Commission. 2024. Standard Contractual Clauses (SCC). https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/standard-contractual-clauses-scc_en.

[16] European Commission. 2025. Report on the first periodic review of the functioning of the adequacy decision on the EU-US Data Privacy Framework. https://commission.europa.eu/document/25695177-8073-4ce3-bf81-eb816dc6b468_en.

[17] Intersoft Consulting. 2025. Art. 45 GDPR Transfers on the basis of an adequacy decision. https://gdpr-info.eu/art-45-gdpr/.

[18] Cookie Database. 2025. Cookie Database. https://cookiedatabase.org/

[19] Atlantic Council. 2022. India's data localization pivot can revamp global digital diplomacy - Atlantic Council. https://www.atlanticcouncil.org/blogs/southasiasource/indias-data-localization-pivot/. (Accessed on 02/20/2023).

[20] Tech Crunch. 2022. *As its data flows woes grow, Google lobbies for quickie fix to EU-US transfers | TechCrunch*. Tech Crunch. (Accessed on 02/20/2023).

[21] Team Cymru. 2023. Whois. whois.cymru.com.

[22] Team Cymru. 2023. Whois. whois.cymru.com.

[23] United Nations Statistics Division. 2024. Standard country or area codes for statistical use (M49). https://unstats.un.org/unsd/methodology/m49/.

[24] EasyList. 2022. EasyList. https://easylist.to/. (Fetched in August 2022.).

[25] Gamero-Garrido, A. and Yu, K. and Shankar, S. V. and Singh, S. K. and Balasubramanian, S. and Wilcox, A. and Choffnes, D. 2025. EU Data Localization Repository. https://github.com/such-in/EU-Data-Localization

[26] GDPRhub. 2021. DSB (Austria) - 2021-0.586.257 (D155.027) - GDPRhub. https://gdprhub.eu/index.php?title=DSB_(Austria)_-_2021-0.586.257_(D155.027)#Further_Resources. (Accessed on 02/20/2023).

[27] L & E Global. 2024. USA: New EU Standard Contractual Clauses – FAQs for U.S. Organisations. https://leglobal.law/2021/07/29/usa-new-eu-standard-contractual-clauses-faqs-for-u-s-organisations/.

[28] Danny S. Guamán, Jose M. Del Alamo, and Julio C. Caiza. 2021. GDPR Compliance Assessment for Cross-Border Personal Data Transfers in Android Apps. *IEEE Access* 9 (2021), 15961–15982. https://doi.org/10.1109/ACCESS.2021.3053130

[29] The White House. 2022. FACT SHEET: President Biden Signs Executive Order to Implement the European Union-U.S. Data Privacy Framework - The White House. https://www.whitehouse.gov/briefing-room/statements-releases/2022/10/07/fact-sheet-president-biden-signs-executive-order-to-implement-the-european-union-u-s-data-privacy-framework/. (Accessed on 02/20/2023).

[30] ICANN. 2024. Registration data lookup tool. https://lookup.icann.org/en.

[31] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. 2018. Tracing Cross Border Web Tracking. In *Proceedings of the Internet Measurement Conference 2018* (Boston, MA, USA) *(IMC '18)*. Association for Computing Machinery, New York, NY, USA, 329–342. https://doi.org/10.1145/3278532.3278561

[32] Ethan Katz-Bassett, John P. John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. 2006. Towards IP Geolocation Using Delay and Topology Measurements. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement* (Rio de Janeriro, Brazil) *(IMC '06)*. Association for Computing Machinery, New York, NY, USA, 71—84. https://doi.org/10.1145/1177080.1177090

[33] Ethan Katz-Bassett and Georgios Smaragdakis. 2021. Seven Years in the Life of Hypergiants' off-Nets. (2021), 516–533. https://doi.org/10.1145/3452296.3472928

[34] Latham and LLP Watkins. 2023. EU-US Data Privacy Framework Goes Live: What Are the Practical Implications? https://www.globalprivacyblog.com/2023/08/eu-us-data-privacy-framework-goes-live-what-are-the-practical-implications/.

[35] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2023. A research-oriented top sites ranking hardened against manipulation - Tranco. https://tranco-list.eu/. (Accessed on 04/10/2024).

[36] Lexology. 2022. Third Time's a Charm? The New EU-US Data Privacy Framework and the US's Pursuit of an EU Adequacy Decision under GDPR - Lexology. https://www.lexology.com/library/detail.aspx?g=b395d1f6-00c6-4081-ae20-91de58e4c109. (Accessed on 02/20/2023).

[37] M. Luckie, B. Huffaker, A. Marder, Z. Bischof, M. Fletcher, and k. claffy. 2021-12. Learning to Extract Geographic Information from Internet Router Hostnames. In *ACM SIGCOMM Conference on emerging Networking EXperiments and Technologies (CoNEXT)*.

[38] Maxmind. 2023. Maxmind Geolocation Data. https://www.maxmind.com/en/geoip2-services-and-databases.

[39] McKinsey. 2023. Data localization and new competitive opportunities | McKinsey | McKinsey. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/localization-of-data-privacy-regulations-creates-competitive-opportunities. (Accessed on 02/20/2023).

[40] Who Tracks Me. 2024. Who Tracks Me. https://www.ghostery.com/whotracksme.

[41] Sebastian Moss. 2024. EU sanctions Rostelecom president, Yandex deputy CEO, and VK Company CEO. https://www.datacenterdynamics.com/en/news/eu-sanctions-rostelecom-president-yandex-deputy-ceo-and-vk-company-ceo/.

[42] Shaoor Munir, Sandra Siby, Umar Iqbal, Steven Englehardt, Zubair Shafiq, and Carmela Troncoso. 2023. Cookiegraph: Understanding and detecting first-party tracking cookies. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 3490–3504.

[43] Yuhong Nan, Xueqiang Wang, Luyi Xing, Xiaojing Liao, Ruoyu Wu, Jianliang Wu, Yifan Zhang, and XiaoFeng Wang. 2023. Are You Spying on Me? Large-Scale Analysis on IoT Data Exposure through Companion Apps. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 6665–6682. https://www.usenix.org/conference/usenixsecurity23/presentation/nan

[44] Noyb. 2020. 101 Complaints on EU-US transfers filed. https://noyb.eu/en/101-complaints-eu-us-transfers-filed. (Accessed on 02/20/2023).

[45] Muhammad Talha Paracha, Balakrishnan Chandrasekara, David Choffnes, and Dave Levin. 2020. A Deeper Look at Web Content Availability and Consistency over HTTP/S. In *2020 Network Traffic Measurement and Analysis Conference (TMA'20)*.

[46] European Parliament. 2020. The CJEU judgment in the Schrems II case. https://www.europarl.europa.eu/RegData/etudes/ATAG/2020/652073/EPRS_ATA(2020)652073_EN.pdf.

[47] Pavlos Sermpezis. 2022. Bias in Internet Measurement Infrastructure. https://labs.ripe.net/author/pavlos_sermpezis/bias-in-internet-measurement-infrastructure/

[48] Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. 2011. IP geolocation databases: unreliable? *SIGCOMM Comput. Commun. Rev.* 41, 2 (apr 2011), 53—56. https://doi.org/10.1145/1971162.1971171

[49] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. 2018. Apps, Trackers, Privacy, and Regulators: A Global Study of the Mobile Tracking Ecosystem. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society. https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_05B-3_Razaghpanah_paper.pdf

[50] RIPE. 2024. RIPE Atlas. https://atlas.ripe.net/.

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

[52] James Saxon and Nick Feamster. 2022. GPS-Based Geolocation of Consumer IP Addresses. In *Passive and Active Measurement: 23rd International Conference, PAM 2022, Virtual Event, March 28–30, 2022, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 122—151. https://doi.org/10.1007/978-3-030-98785-5_6

[53] Pavlos Sermpezis, Lars Prehn, Sofia Kostoglou, Marcel Flores, Athena Vakali, and Emile Aben. 2023. Bias in Internet Measurement Platforms. In *2023 7th Network Traffic Measurement and Analysis Conference (TMA)*. 1–10. https://doi.org/10.23919/TMA58422.2023.10198985

[54] Amazon Web Services. 2024. AWS IP Address Ranges. https://ip-ranges.amazonaws.com/ip-ranges.json Accessed: February 28, 2025.
[55] SimilarWeb. 2022. Top Sites. https://www.similarweb.com/top-websites/germany/. (Data for July 2022.).
[56] Ao-Jan Su, David R. Choffnes, Aleksandar Kuzmanovic, and Fabián E. Bustamante. 2006. Drafting behind Akamai (travelocity-based detouring). In *Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications* (Pisa, Italy) *(SIGCOMM '06)*. Association for Computing Machinery, New York, NY, USA, 435—446. https://doi.org/10.1145/1159913.1159962
[57] Tessian. 2022. *30 Biggest GDPR Fines To-Date | Latest GDPR Fines | Updated 2022 | Tessian.* Thessian. (Accessed on 02/20/2023).
[58] European Union. 2015. Judgment of the Court (Grand Chamber) of 6 October 2015. Maximillian Schrems v Data Protection Commissioner. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A62014CJ0362.
[59] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Beyond the Front Page:Measuring Third Party Dynamics in the Field. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1275–1286. https://doi.org/10.1145/3366423.3380203
[60] Verizon. 2022. Monthly IP Latency Data Verizon Enterprise Solutions. https://www.verizon.com/business/terms/latency/. (Data for August 2022.).
[61] WonderNetwork. 2022. Global Ping Statistics - WonderNetwork. https://wondernetwork.com/pings. (Data for November 2022.).
[62] Alexander R Zheutlin, Joshua D Niforatos, and Jeremy B Sussman. 2021. Data-tracking on government, non-profit, and commercial health-related websites. *Journal of general internal medicine* (2021), 1–3.

A LATENCY DISTRIBUTIONS

In Fig. 9, we include the latency distribution from our measurements. We include both source- and destination-based mesurements, both of which we use (sequentially) to confirm the location of servers in non-adequate countries. The series in the charts correspond to the latency observed to all candidate non-adequate servers as well as those that we still label as non-adequate after the application of speed of light constraints. As expected, the subset of servers that we confirm as being located in non-adequate countries tend to have higher latencies than the larger group of initial candidate servers.

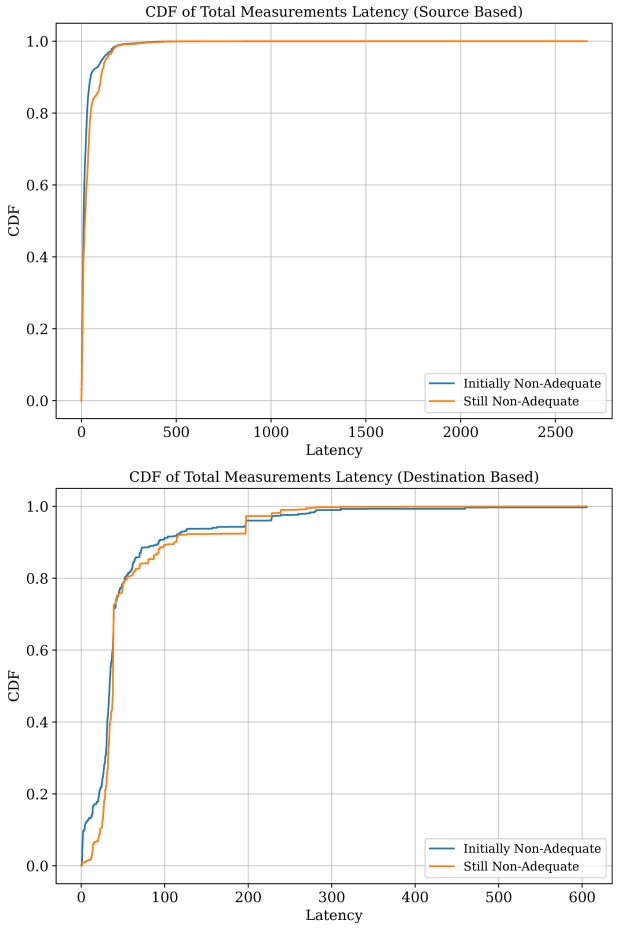


Figure 9: CDF of latency observed in source- and destination-based measurements.

B ETHICS

Our study does not collect personal information of any kind and does not qualify as human subjects research. We collect data only from public, very popular web sites and the resources that they load. We do not log in to any sites. We use a separate browser that is routed through a proxy (not the user's own browser). Thus we do not have access to any user's browsing history, data locally stored in any user's device, any device/identifying information, nor any other private data. The only exception is our limited experiment to confirm the accuracy of BrightData's stated location of proxies; there, we only record the /24 subnet to which the device running the proxy is connected. In sum, our study is an investigation of the public web and does not pose significant ethical concerns. The scale of our data collection, for all types of information we gather, is unlikely to impact any services of the major websites we study, and there are meaningful benefits in understanding compliance with online privacy regulations.