A Framework for Assessing Adherence with Data Localization Policies

Alexander Gamero-Garrido, Kicho Yu, Sumukh Vasisht Shankar, Sindhya Balasubramanian, Alexander Wilcox, David Choffnes

1 NETWORK & COUNTRY SAMPLE

Our method ensures that we launch both browser-based and network measurements from the same network and in the same country. 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 [1], a large-scale Internet measurement platform with very dense deployment in the EU; and BrightData, a large-scale proxy service. [2] We 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.

While RIPE regularly publishes a list of its active probes [1], 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.

2 IDENTIFYING RELEVANT DOMAINS & TRACKERS

In this section, we describe our identification of relevant, popular domains in each EU country, step 1 in Fig. 1.

2.1 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 [10]. (Alternatives, including most prominently the Tranco list, are inadequate for our purposes as they provide a single ranking for the global web, rather than per-country lists [5].) 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.

2.2 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. [3]. 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. [4] Domains that are not first parties are then labeled as third parties.

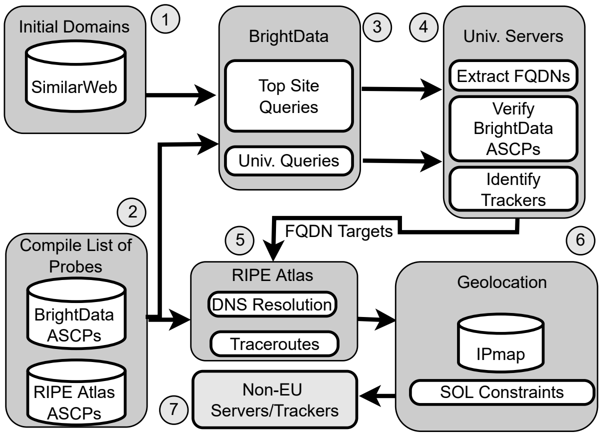

Figure 1: Process diagram summarizing our methodology.

2.3 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 (§ 2.2). 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.

[1] atlas.ripe.net

[2] brightdata.com

[3] whois.cymru.com

[4] https://www.caida.org/archive/as2org/

Pre-Print YYYY(X), 1–4

https://doi.org/XXXXXXX.XXXXXXX

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. 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 able to infer their data localization compliance with our method.

2.4 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 [7], 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.

2.5 Web Crawls Through BrightData

We "browse" all popular sites in each country and record their response. 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, we also record the web contents and cookies, and plan on releasing them with the rest of our data and code upon paper acceptance.

2.6 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. To label a domain as a tracking site, we use a three-step approach applied to the domains found in the Selenium DNS requests (§ 2.5). First, we intersect the domains with known trackers from the well-established list, EasyList (easylist.to). After manually inspecting the 256 third party domains (§ 2.2) 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 [5]; 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).

3 SERVER GEOLOCATION

This section covers our method to locate servers in EU nations.

3.1 Source-Based Measurements

We first obtain a preliminary assessment of where the server is located using RIPE IPmap. 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. Then, we identify candidate servers that may be located in non-adequate contries since both the traceroute latency and IPMap support that inferred location.

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 [11]. Since Verizon does not publish latency data between Latin America and the EU, we rely on wondernetwork.com/pings [12] 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). 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

[5] https://badmojr.gitlab.io/1hosts/

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.

3.2 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. We then use speed of light (subsequently denoted by c) constraints to discard likely erroneous geolocation inferences by IPMap. Our goal is to minimize the rate of false positives, or the number of servers inferred to be in non-adequate countries that actually are in an adequate country.

We launch traceroutes from RIPE Atlas probes located in the same non-adequate country where the server was inferred to be located by IPMap. 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 § 3.1. 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 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 § 3.1, 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 minimize the rate of false positives 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 $2c/3$ [4], or infrastructure delays, such as queue buildups on routers, our upper bound for observed speed is $4c/9$ [3] or approx. $133 km/ms$; 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.

3.3 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). Hostnames obtained from rDNS are often, but not always, useful in geolocating IP infrastructure [6], 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 [6], 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 [2]).

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. We are left with 396 measurements (to 247 IP addresses) where all available evidence suggests that the server responding to EU requests is located in a non-adequate country.

4 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 our 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 removing the last octet. We then compare the country and AS claimed by BrightData with those identified by geolocation database Maxmind. 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 5 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 5 exceptions include 2 where the country does not match (but the AS does), 2 where the

AS does not match (but the country does), and 1 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 [8, 9]. 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.

REFERENCES

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

[2] 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).

[3] 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, 71âĂŞ84. https://doi.org/10.1145/1177080.1177090

[4] 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

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

[6] 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)*.

[7] 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)*.

[8] 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), 53âĂŞ56. https://doi.org/10.1145/1971162.1971171

[9] 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 28âĂŞ30, 2022, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 122âĂŞ151. https://doi.org/10.1007/978-3-030-98785-5_6

[10] SimilarWeb. 2022. Top Sites. https://www.similarweb.com/top-websites/germany/. (Data for August 2022.).

[11] Verizon. 2022. Monthly IP Latency Data Verizon Enterprise Solutions. https://www.verizon.com/business/terms/latency/. (Data for August 2022.).

[12] WonderNetwork. 2022. Global Ping Statistics - WonderNetwork. https://wondernetwork.com/pings. (Data for November 2022.).