Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, Nikolaos Laoutaris

Tracing Cross Border Web Tracking

Iordanou, C., Smaragdakis, G., Poese, I., & Laoutaris, N. (2018). Tracing Cross Border Web Tracking. Proceedings of the Internet Measurement Conference 2018 - IMC ’18. Presented at the Internet Measurement Conference 2018. https://doi.org/10.1145/3278532.3278561

Tracing Cross Border Web Tracking

Costas Iordanou
TU Berlin / UC3M

Georgios Smaragdakis
TU Berlin

Ingmar Poese
BENOCS

Nikolaos Laoutaris
Data Transparency Lab / Eurecat

ABSTRACT

A tracking flow is a flow between an end user and a Web tracking service. We develop an extensive measurement methodology for quantifying at scale the amount of tracking flows that cross data protection borders, be it national or international, such as the EU28 border within which the General Data Protection Regulation (GDPR) applies. Our methodology uses a browser extension to fully render advertising and tracking code, various lists and heuristics to extract well known trackers, passive DNS replication to get all the IP ranges of trackers, and state-of-the art geolocation. We employ our methodology on a dataset from 350 real users of the browser extension over a period of more than four months, and then generalize our results by analyzing billions of web tracking flows from more than 60 million broadband and mobile users from 4 large European ISPs. We show that the majority of tracking flows cross national borders in Europe but, unlike popular belief, are pretty well confined within the larger GDPR jurisdiction. Simple DNS redirection and PoP mirroring can increase national confinement while sealing almost all tracking flows within Europe. Last, we show that cross boarder tracking is prevalent even in sensitive and hence protected data categories and groups including health, sexual orientation, minors, and others.

1 INTRODUCTION

Online advertising, including bahavioral targeting over the Real Time Bidding protocol (RTB) [62], fuels [26] most of the free services of the web. In its principle, the concept of targeted (or personalized) advertising is benign: products and services offered to consumers that they truly care about. It is in its implementation and actual use when controversies arise. For example, tracking should respect fundamental data protection rights of people, such as their desire to opt-out, and should keep clear from sensitive personal data categories, such as health, political beliefs, religion or sexual orientation. One of the most important changes on how to process and store personal data is the European Union General Data Protection Regulation (GDPR) [5]. GDPR offers protection to European citizens across a wide range of privacy threats, including tracking on sensitive categories such as those mentioned above. Now that Europe's new data protection law is in place (implementation date of the GDPR across the European Union was on May 25, 2018; the regulation entered into force on May 24, 2016), the next challenge becomes implementing it in practice. GDPR has provisions that include steep fines reaching up to 4% of worldwide turnover or 20 million euros, whichever is higher, for any company found in violation. Monitoring the effectiveness of the law, investigating complaints, and prosecuting violators can only be carried out based on sound factual data. The measurement community, therefore, has an important role to play in developing the necessary new methodologies and in collecting data for GDPR related topics and investigations.

A fast growing body of literature already exists around topics such as "What information is leaking while users navigate the web with fixed [28–30, 35, 41, 43, 44, 51, 58, 61] or mobile devices?" [42, 52, 53, 60], "Who is collecting it?" [29, 52, 58], "How is it being collected?" [27, 47, 57], "What is its financial worth?" [48, 49], "Which are the potential hazards for citizens?" [45], *etc.* (see Sect. 8 for more related work). An area, however, that has received relatively small attention has to do with the geographical aspects of tracking, including questions such as: Where is the back end of a tracker?, How far does a tracking flow go?, Which borders does it cross?, What can be done to contain tracking within a certain data protection jurisdiction?

Extracting the geographical footprint of trackers and tracking flows is difficult for a number of reasons: It requires having access to real tracking flows originating from real users and terminating at dynamically bound trackers. The obtained sample needs to be representative, unbiased, and complete in terms of coverage. The obtained measurements need to be precise, especially in terms of geolocation accuracy.

Our contribution: In this paper, we develop a novel measurement methodology for mapping the geographic characteristics of tracking flows at scale. Our methodology is hybrid in nature – it is using fully rendered webpages and executed tracking code to detect tracking flows. For this we use a population of test users from the CrowdFlower platform [4] who have installed our browser extension. With trackers identified, we then look them up in large NetFlow datasets from entire ISPs. Therefore, our browser extension is adding *precision*, and our lookup step, *scale.* An important intermediate step has to do with guaranteeing the *completeness* of the lookup, *i.e.,* that we have identified all the IPs of a tracker, and that we confirm that these IPs are dedicated for tracking. For this we utilize DNS databases (archived passive DNS records, see Sect. 3.3).

In summary, our methodology manages (i) to double the amount of tracking flows detected compared to previous simpler approaches, (ii) improve their geolocation accuracy, and (iii) monitor the tracking ecosystem continuously for a time period of more than four months capturing any possible temporal variations.

Our findings: By applying our methodology on data from 350 CrowdFlower users and NetFlow data from 60M ISP subscribers, we show that:

- Most tracking flows, typically around 90%, originating at users within EU28 terminate at tracking servers hosted within EU28. This result contrasts popular belief, as well as recent studies, claiming that most tracking of European citizens

is conducted by trackers physically located outside Europe. The discrepancy owes to geolocation accuracy, among other reasons.
- Confinement within national borders is much lower: peaking at less than 70% in the best case and becoming single digit for small countries. There exists a correlation between the density level of IT infrastructure of a country, mostly in terms of datacenters, and the confinement of tracking flows within its borders.

Subsequently, we turn our attention on what can be done to improve the locality of tracking flows. We consider two mechanisms: DNS redirection and PoP mirroring.

- With a more thoughtful DNS redirection on behalf of the tracking domains administrators, the overall confinement percentage can be improved at both, country and continent level, at a minimal financial cost for the tracking domains.
- Applying PoP mirroring over popular public clouds also improves the confinement percentage within the GDPR region, but when applied on top of locality-improving DNS redirection, the improvement at national level is rather marginal for all but a few countries.

Last, we look at sensitive personal data by tracing tracking flows induced by websites involving sensitive categories, such as, ethnicity and sexual orientation.

- We show that despite the threat of steep fines under GDPR, around 3% of the total tracking flows identified, relate to protected data categories.
- The percentage of such flows crossing borders appears to be similar with that for general tracking traffic.

2 BACKGROUND

Before delving into technical details, we outline the current legislative and operational setting that motivate our study, and highlight the challenges we want to overcome.

2.1 Why location matters?

As with most things, location matters also with data protection. This may seem counter-intuitive since GDPR only requires that an online service access a European citizen's data to hold it accountable independently of the location of its legal or technical base. Thus a company incorporated in the US with its servers in, for example, Singapore can still get fined if it fails to conform to GDPR requirements while processing data of European citizens. Then why is it important to know whether a tracking flow crosses the EU28 borders? The answer is – *investigation & enforcement*. Indeed data protection complaints can be investigated in greater depth when a Data Protection Authorities (DPA) can be granted legal access to the tracking backend. This is far easier done when the tracking end point is within EU28 borders.[1]

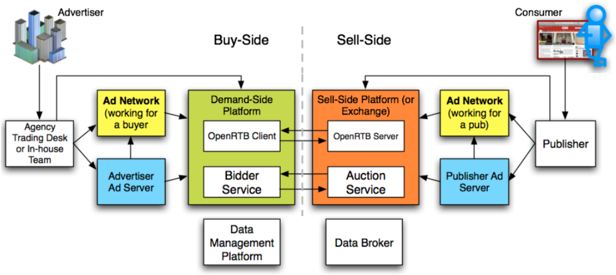


Figure 1: High-level communication between parties in the Open RTB Ecosystem [25] (courtesy: Interactive Advertising Bureau).

But what about national borders? These are important for *jurisdiction reasons*. Although GDPR is the common data protection law of all EU28 countries, its implementation is left to the corresponding national DPAs. The national DPA is responsible for the handling of a complaint of the citizen or legal entity filling the complaint. Therefore, it is important to know how many tracking flows cross national borders and where the tracking servers are physically hosted.

Last but not least, other pieces of legislation exist that may impact on tracking (*e.g.,* security related, protection of minors, data or server logs storage duration *etc.*), which only have a national scope. For these cases it is also important to know whether a tracking flow stays within national borders.

2.2 Online tracking over RTB

Figure 1 depicts a high level block diagram of the different entities involved in targeted advertising over real-time bidding (RTB), one of the main advertising and marketing applications that require tracking end users across different publisher websites.[2] For a detailed description of the different entities, the reader is referred to [62]. Virtually all the entities depicted in the diagram may be present with advertising and/or tracking code at the publishers website and thus, be rendered by a consumer's browser while visiting the publisher. The execution of such code induces tracking flows between the consumer and the corresponding entity. Of course there are additional flows related to tracking that are exchanged directly between the entities without going through the end user's device. In this paper, we report only upon the directly visible tracking flows, *i.e.,* those that involve execution of tracking code at the consumer's browser.

2.3 Challenges

Tracing the geographic aspects of tracking flows has received relatively little attention. This is not surprising given the involved technical challenges.

Challenge 1 - Collecting real tracking flows: Existing work has gone mostly into quantifying and cataloging the trackers found present in different publisher websites [29, 35, 36, 41, 58]. This is

[1] Notice that terminating a tracking flow within Europe does not guaranteed that the personal data of citizens have not flown outside the continent. Once collected by a tracker the data can be moved in any place in the world in a variety of means. Having the terminating end-point within Europe, however, is important since it allows a more thorough investigation to access the end-point and verify what data have been collected and where else they have been transmitted. Data can also cross in and out of Europe multiple times while in transit due to IP routing. In the process eavesdroppers can take a look at them. We don't consider such matters since they are subject to different laws about telecommunications and surveillance.

[2] Open RTB [14] is the industrial standard of Interactive Advertising Bureau (IAB).

already challenging since identifying tracking code requires full rendering of publisher webpages. In our case things are even more difficult since rendering webpages through automated crawlers is not enough – we need real users, with real credentials and web browsing history, at different locations to capture the full spatial aspects of tracking flows. Releasing measurement code to real users is difficult to scale, while passive network logs, *e.g.,* NetFlow [32] or sFlow [55], are at a much higher level that makes identifying tracking flows difficult, either because there is no payload (NetFlow) or there is partial or encrypted payload (sFlow).

Challenge 2 - Completeness of measurement: If one attempts to combine the precision of full rendering via dedicated measurement code with the scale of passively collected network logs, they will eventually run into issues of completeness such as: Are there any additional tracking domains other than the ones seen by the measurement code? Which are the IPs associated with these domains? Are there any additional IPs associated with these domains that were not returned to the real users?

Challenge 3 - Precision of analysis: Collecting complete measurements is only a first step. Next, the analysis has to be conducted with care. For our study, *accurate IP geolocation* is key to deriving reliable results and conclusions. It is well known that infrastructure IPs, such as servers [31, 56] and routers [34, 37, 39, 50], are prone to imprecise geolocation. It is also important to investigate if the tracking IPs *are dedicated to tracking*, or are shared with other services and domains. It is also important to identify the time *period* that a specific IP is associated with a tracking service in order to remove noise from dynamic use of IPs.

3 METHODOLOGY

To address Challenge 1, we present in this section the design and implementation of a browser extension for identifying tracking flows triggered by real users' actions and data. We also outline methodologies for improving the completeness and the precision of our measurements, thereby addressing Challenges 2 and 3, respectively.

3.1 Our browser extension

To identify as many ad and tracking related domains, and their associated IPs, as possible, requires collecting visits of real users to websites (first-party request/domain) that embed such services. User state information, such as their browsing history, cookies, exact location, time of visit, *etc.,* impact on the behavior of tracking, *e.g.,* through winning bids and tracking connections opened. Furthermore, using real users' browsers has the additional advantage of capturing the interaction of the user with the elements of the webpage, which itself alone can lead to the launching of additional tracking requests. Finally, it is important to monitor requests from as many geographic locations as possible, which becomes easier when having a real user base across the globe. Many of the above are impossible to achieve using scripted crawlers launched from few measurement locations that do not correspond to usual residential broadband networks and real user behavior.

To address all of the above, we have developed and distributed a browser extension for Google Chrome. The extension is used for a related measurement project about targeted advertising detection. In the process, however, we obtain valuable data for this study as well. Specifically, we can identify and monitor all outgoing third-party requests, *i.e.,* requests towards domains apart from the one that the user is actually visiting during their normal browsing sessions. For each outgoing third-party request, our extension maps the associated server IP as observed in the corresponding response header. Since we operate in the users' browser, we only focus on the final server that serves the third-party requests. The browser API does not report on JavaScript or DNS redirections. However, it reports the final IP that serves a request.

Table 1: The real users dataset statistics.

# Users	# 1st party Domains	# 1st party Requests	# 3rd party Domains	# 3rd party Requests
350	5,693	76,507	19,298	7,172,752

The small number of published reports based on measurements from real users is indicative of the involved difficulties related with such studies. To the best of our knowledge, apart from Razaghpanah *et al.* [52] for mobile apps, this is the first time that a crowdsourced approach is utilized to report on geographic aspects of tracking using real users' data.

We recruited users from the CrowdFlower platform [4]. We excluded users having ad-blocking extensions installed on their browser, such as, AdBlockPlus, Ghostery, *etc.* In total, 350 users have installed the browser extension and contributed data to our study in a time window of more than four months, from Sep. 1, 2017 to mid-Jan., 2018. The number of unique visited websites is more than 76K and the total third-party requests logged exceeds 7.1M over more than 19K third-party domains. For a detailed summary, see Table 1. The collected dataset includes the user's country, the first-party visited domain, the third-party contacted URL and the associated IP.

Ethical considerations: As we have previously stated, all the users in this study were recruited from the CrowdFlower platform [4]. All the users were informed in detail about what data we collect and gave their explicit consent *before* installing the extension. Users could choose at any point to opt-out of the experiment by simply uninstalling the browser extension. This would stop any data transfer to our servers. Regarding already collected data, although we can delete any part of it, it's impossible to identify specific users since we do not store any unique identifier on users. For example, we did not keep logs of actual IPs, but only the geolocated regions. We also took additional measures to protect the identity of the user, namely, we only collected domain names instead of full URLs. Thus, we avoided inadvertently collecting the full browsing history of a user, or storing identity information that may appear on URLs. Obviously we refrained from asking or collecting any personally identifiable information such as the real name of the users, emails, addresses, *etc.* All users were compensated through the platform for keeping the browser extension running for an amount of time.

3.2 Identifying trackers

In this section we explain our methodology for identifying whether a third-party request is actually a tracking flow or just some other

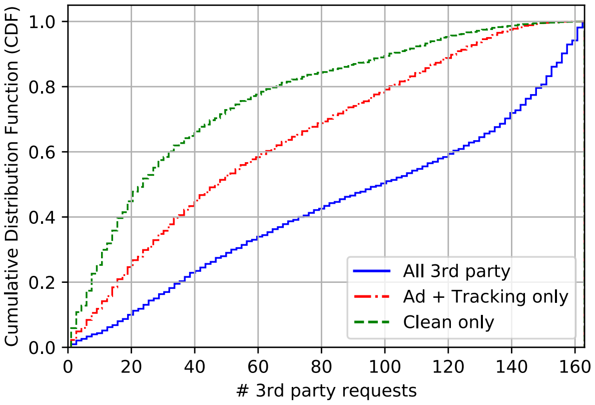


Figure 2: The number of 3rd party requests per website based on data collected from real users browsers. *"Clean only"* (top) depicts the flows related to other activities such as, live chat services, commenting services, etc. *"Ad + Tracking only"* (middle) depicts the flows related to ad and tracking, and finally *"All 3rd party"* (bottom) depicts the CDF of all the flows.

Table 2: Comparison between the AdBlockPlus lists (easylist, easyprivacy - top row) and the semi-manual classification (bottom row) to identify ad and tracking related third-party requests.

	# FQDN	# TLD	# Unique Requests	# Total Requests
AdBlockPlus Lists	6,259	1,863	539,293	2,446,460
Semi-automatic	3,620	879	453,457	1,964,408
Total	9,879	2,742	992,750	4,410,868

type of service (*i.e.,* voice chat, commenting services, *etc.*). Currently, the most common solution is to use a block-list. The most popular lists for detecting ad- and tracking-related requests are the *"easylist"* and the *"easyprivacy"* [7] list, respectively. The issue with the above two lists is that they are constructed and used for *blocking* third-party requests from web browser extensions, such as, the AdBlockPlus [1] and Ghostery [9]. By blocking a tracking flow early, they do not allow any additional tracking code to be executed, which in turn may open additional connections and thereby reveal additional tracking requests that do not match any rules or domains in the above two lists.

To overcome the above limitation, we first use the above two lists (easylist and easyprivacy) to classify all the third-party flows that we collect either as tracking or not. This produces a list of tracking flows (LTF) that includes all third-party requests that the two filtering lists identify as ad or tracking related requests and a list of non-tracking flows (NTF). As a second step, we use the list of LTF to classify additional third-party requests. We examine if the referrer field of the remaining non-tracking flows in the NTF list includes any URL already detected in the LTF list and also if the URL string includes any arguments. Note that argument parsing using the URL is a widely used technique for passing information between tracking domains. If a non-tracking flow satisfies both requirements we then classify it also as a tracking flow. Note that the execution of additional requests using third-party code (JavaScript) embedded directly into the first-party context populates the referrer field of the request with the first-party URL. Nevertheless, most of this cases are requests towards well known ad networks to initialize the rendering process of the available ad slots within the first-party webpage, such as, googlesyndication.com.

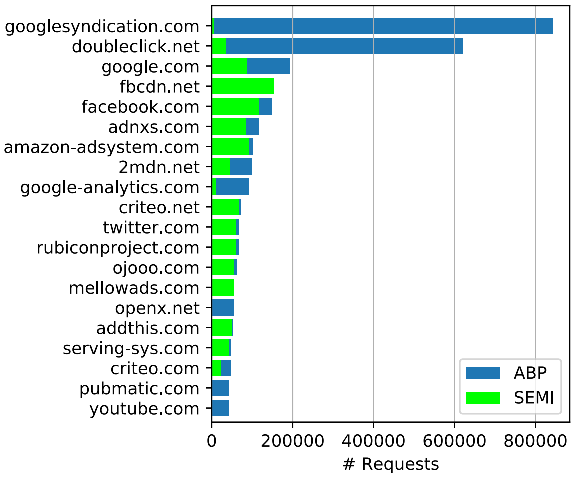


Figure 3: The top 20 TLD of ad + tracking domains based on requests counts in the real users dataset.

Finally, for the remaining non-tracking flows, we also classify third-party requests as tracking flows when the request URL include arguments and also the URL string include some widely used keywords related to web tracking and advertising, such as, "usermatch", "rtb", "cookiesync", *etc.* Note that we build the list of keywords empirically.

Table 2 presents the third-party requests classification results. Using the two AdBlockPlus lists (easylist, easyprivacy), we manage to classify a total of 2.4M third-party requests as tracking flows (Table 2 - Row 1). In total, we have more than 500K unique URLs towards 1.8K top level domains (TLD). Using our semi-automatic classification (Table 2 - Row 2), we manage to classify an additional 1.9M third-party requests as tracking flows from more than 400K unique URLs and a total of 879 top-level domains.

In Fig. 2 we plot the CDF of the tracking and non-tracking flows that we detect in each website in our dataset. The top (dashed) line depicts the CDF of the non-tracking flows and the middle (dot-dashed) line the tracking flows. Finally, the bottom (solid) line depicts the total of all requests that we observe including both tracking and non-tracking flows from within each website. The main takeaway from Fig. 2 is that on average, most of the third-party requests are ad and tracking related flows.

Finally, Fig. 3 lists the top 20 TLDs of the tracking flows that we detect in our dataset. "ABP" denotes the number of tracking flows detected using the AdBlockPlus lists and "SEMI" denotes the ones detected using the semi-automatic classification. We observe

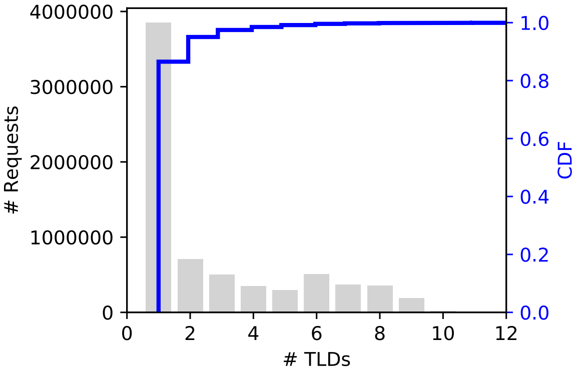


Figure 4: The CDF of number of domains detected behind each IP (right y-axis) and the total number of request (left y-axis) observed in the real users dataset.

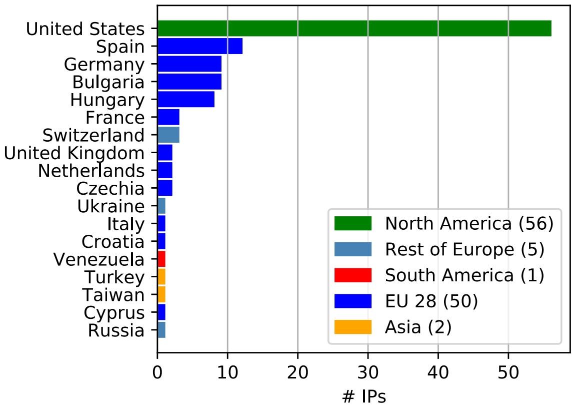


Figure 5: The number of IPs that host more than 10 ad + tracking domains and their corresponding geolocation.

that most of the additional tracking flows detected by the semi-automatic methodology involves domains belonging to ad networks, mostly triggered by the (potentially blocked) ad related initial third-party request and constituently not detectable by ABP.

3.3 Collecting tracker IPs

For each third-party domain, we collect all the associated IPs that were returned to users who successfully established a connection. Real users from all over the world participated in our four-month experiment. In total, we collected 28,939 tracking IPs. More than 97% of them were IPv4.

Furthermore, to address Challenge 2, *i.e.,* to improve completeness of our measurement, we took some additional steps. First, we utilized passive DNS replication (pDNS) [63], a method that collects DNS data from production networks and stores it in a database for later reference. In this work we rely on Robtex implementation of pDNS [22]. These databases provide info on (i) forward DNS records, *i.e.,* the IPs associated with a given domain as well as the starting and the end of the time period of this association, and, (ii) reverse DNS records, that map an IP for a given time period to the domains that were served by this IP. For the duration of the experiment, we identified only 806 additional IPs (*i.e.,* small 2.78% increase on the number of IPs, mainly IPv4 (60%) that served the tracking domains but could not identify from the logs of the real users. We also annotated the active periods for the pair domain-IP based on the starting and end active time in the database.

Next, we investigate if other services/domains share the same IP. In Fig. 4, we plot the histogram and the CDF for the number of TLDs served by an identified tracking IP weighted by the number of requests. Around 85% of the requests served by IPs serve only one TLD. This is to be expected as tracking services would like to sustain a good performance and thus, dedicate the IP for this service. Delays may reduce revenue, and if the tracker is involved in RTB, it is important to guarantee a short round trip time with the user, as the bidding time is typically in the order of 100 msec [13]. In the same figure also shows that the fraction of IPs that serve more than one domain is less than 2%. A closer investigation shows that the other TLDs usually belong to the same organization, and they are tracking related domains as well (*e.g.,* in the case of Google, doubleclick.net and googlesyndication.com). Thus, measuring the flows that involve the identified tracking IPs, for the time period that the pair tracking domain and tracking IP is valid, will give us a good estimation of the tracking flows.

Nevertheless, there are a few IPs (114 in total) – about half of them in the USA and in EU28 (see Fig. 5) – that serve a large number of domains – 10 or more. A closer investigation showed that these IPs are used for ad related activities, such as, ad exchange points, RTB auctions or cookie-syncing as they serve a large number of domains related to the advertisement and tracking industry.

3.4 Geolocating web tracker IPs

To address Challenge 3, we geolocate the ad and tracking related IPs as accurately as possible in order to minimize artifacts that can bias our analysis. It is well reported that commercial geolocation databases are unreliable when it comes to geolocating network infrastructure [31, 34, 37, 39, 50, 56]. This is expected as the commercial interest of these databases is to geolocate the end user accurately – the customers of such databases are enterprises that want to geolocate their visitors/clients. Several existing studies have shown that commercial databases, such as MaxMind [16], are particularly bad for geolocating web servers [31, 56]. For example, in the case of Google, MaxMind typically geolocates a Google IP to Mountain View, the headquarters of Google and not to the real physical location of the server, which can be at any Google datacenter, at a peering facility (edge point of presence), or even inside an ISP (edge cache) [12, 31, 59].

A number of active techniques have been developed for improving the IP geolocation accuracy for the server infrastructures [31]. RIPE has incorporated these techniques in a single publicly available tool called RIPE IPmap [21]. IPmap uses a large global installation of more than 11K active measurement probes, namely RIPE Atlas [20], to perform active measurements in order to geolocate an IP. The

Table 3: Pair-wise agreement across geolocation tools.

Service	ip-api		MaxMind		RIPE IPmap	
	Country	Cont.	Country	Cont.	Country	Cont,
ip-api	100%		96.13%	99.15%	53.24%	65.62%
MaxMind	96.13%	99.15%	100%		53.4%	64.96%
RIPE IPmap	53.24%	65.62%	53.4%	64.96%	100%	

Table 4: Wrong geolocated IPs/Requests using MaxMind database for Google, Amazon and Facebook ad and tracking domains.

	# IPs	Wrong Country	Wrong Cont.	# Requests	Wrong Country	Wrong Cont.
Google Ads + Tracking	4,873	2,822 57.91%	2,099 43.07%	1,941,301	1,231,298 63.43%	1,157,910 59.65%
Amazon Ads + Tracking	3,306	1,951 59.01%	1,948 58.92%	165,181	53,434 32.35%	53,109 32.15%
Facebook Ads + Tracking	646	292 45.20%	191 29.57%	67,805	8,181 12.06%	5,279 7.79%

footprint of the RIPE Atlas probes is particularly dense in Europe (more than 5K probes) thus, in Europe the accuracy is expected to be high, especially at country level, which suffices for our study. RIPE Atlas has also a large footprint in the US, with more than 1K probes thus, using IPmap we can accurately distinguish if a server is in Europe or in the US. For every IP geolocation request, more than 100 RIPE Atlas probes are assigned to perform active measurements. After the geolocation process is finished, each probe replies with an estimation of the physical location of the target (server in our study) at the city, country, and continent level. We noticed that, across all our measurements, the replies from the involved probes agree on the continent, and also with a majority of above 90% on the country. We also noticed that the disagreement on the country level (less than 10%) occurs around the borders of neighboring countries. For our analysis, we do a majority voting and we keep the most popular estimation. To further evaluate the accuracy of RIPE IPmap, we geolocated the IP ranges of two large content providers, Amazon AWS [2] and Microsoft Azure [17], that made the location of the servers in these ranges publicly available. Our analysis about the active IPs that replied to our requests showed that RIPE IPmap accurately geolocate the server IPs at both country (99.58%) and continent level (100%) for the above two cloud services.

In Table 3 we compare the pair-wise agreement on the country and continent, across geolocation tools, namely, (i) IP-API free geolocation tool [15], (ii) MaxMind [16], and (iii) RIPE IPmap [21], for the tracking IPs we inferred with the browser extension (including the additional IPs we found with forward DNS). The overlap between IP-API and MaxMind is very high, more than 96% on the country level and 99% on the continent level. However, both disagree when compared with the IPmap. About half of the IPs are mapped to a different country and approximately a third of the IPs are mapped to a different continent. This is an indication that using MaxMind or IP-API would yield incorrect geolocation in our analysis, since one of the end points of all our flows is always a backend infrastructure server.

To further investigate the impact of the MaxMind database as opposed to RIPE IPmap, we concentrate on three large ad + tracking provider, namely, Google Ads+Tracking IPs, Amazon Ads+Tracking IPs[3] and Facebook Ads+Tracking IPs. In Table 4 it is clear that about half of the IPs of these major providers are mis-geolocated to the wrong country, and anywhere between 30%-60% are mis-geolocated to the wrong continent.

4 QUANTIFYING BORDER CROSSING

In this section, we present our measurement results on the amount of tracking flows crossing different national and international borders. All the results of this section are based on measurements obtained with our browser extension and recruited users. Later in Sect. 7, we present corresponding results from four large ISPs with more than 60 million users.

Figure 6 shows the percentages of tracking flows exchanged between continents (or geographic regions like EU28). The thickness of the Sankey diagram is proportional to the amount of measurements that we have from each region. We see that *most tracking originating at users within EU28 terminates at tracking servers within EU28.* The actual percentage is 84.9% as shown in the more detailed Fig. 7(b). This result contrasts popular belief, as well as recent reports [52] claiming that most tracking of European citizens is conducted by trackers physically located outside Europe. The discrepancy is explained by the different IP geolocation methods used (see Sect. 3.4 for details) but also owes to other reasons. For example, in the case of [52] the variations are also due to difference in the platforms in use (Mobile *vs.* Desktop in our case) and the variation between the two platforms (mobile apps *vs.* web browsing), see Sect. 8 for more details.

Unlike EU28 that exhibits high confinement of tracking flows within the continent, our second larger user base in South America sees most of its tracking flows (95%) leaking out of the continent and into North America (90%). Since we mainly focused on recruiting European users for our study, the other continents shown in the diagram have small user bases and therefore the confinements ratios are not easy to read from the diagram. The actual numbers are Africa 2.11% (22), Asia 16.39% (20), Rest of Europe 12.94% (23), South America 4.42% (86), North America 86.83% (16), confinement percentage (number of users), respectively.

Overall, we see that EU28 and North America host most of the tracking backends, 51% and 40% of all traffic flow terminations, respectively. Other countries, with large IT infrastructure/server hosting receive a disproportionally high number of flows compared with the users in our dataset, *e.g.,* Ireland (3.4%), Switzerland (2%), France (6%), Russia (1.5%).[4]

4.1 EU28 GDPR jurisdiction

In the remainder of the paper, we focus on tracking of users in EU28, where we have our largest user base (183 users). Figure 7(a) shows the percentage of tracking flows that terminate in different continents for users within EU28 under MaxMind geolocation. Figure 7(b) shows the same percentages under RIPE IPmap geolocation, and the difference in numbers is astonishing. In fact, this single property of the methodology - the method used for IP geolocation can flip the qualitative takeaway of the result. Under MaxMind one

[3] Amazon uses different IP addresses for such activities not included within the AWS IP ranges

[4] Results not shown in the diagram.

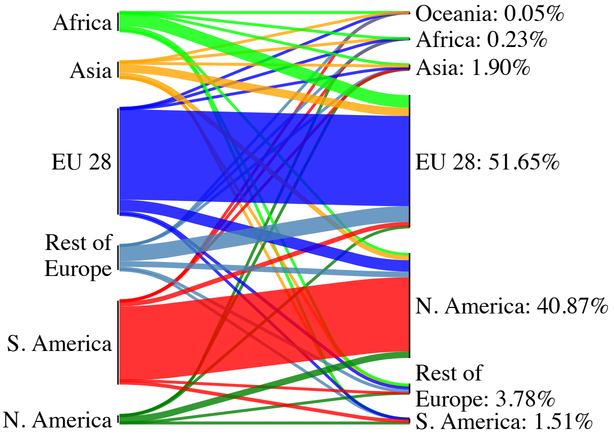
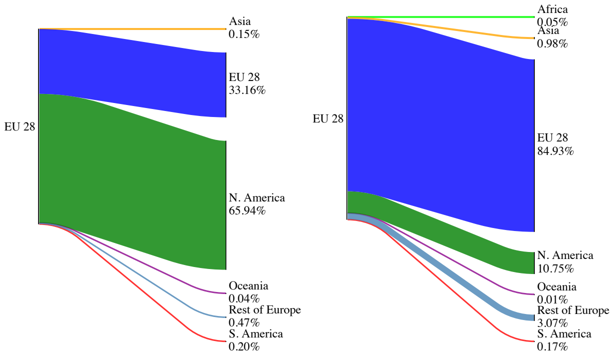


Figure 6: The flow of ad + tracking domains between continents using the RIPE IPmap geolocation service.



(a) MaxMind

(b) RIPE IPmap

Figure 7: The flow of ad and tracking domains between continents from EU28 countries using the (a) MaxMind and (b) RIPE IPmap geolocation services.

concludes that most European tracking flows leak towards North America, whereas under RIPE IPmap, they remain confined within Europe. As explained in Sect. 3.4, RIPE IPmap is way more accurate for the problem at hand and therefore we conclude that *most tracking flows affecting European citizens terminate within GDPR's legal jurisdiction*. The only sizeable percentage leaking outside Europe is towards North America (10% of European tracking flows). Another 3% goes to neighboring non-EU28 European countries, such as, Switzerland and Russia.

4.2 National jurisdiction

Figure 8 is a Sankey diagram for the origin-destination of tracking flows originating in EU28 countries, where we have users in our dataset (the thickness of a flow is proportional to the user base in each country on the left column). We observe different levels of national confinement. The UK leads with the highest confinement of 58.4% within its borders. Spain follows with a confinement of 33.1%. Smaller counties like Greece, Romania, and Cyprus have lower confinements, 6.77%, 5.1%, and 1,16% respectively. From this data, there appears to be a positive correlation between the size of a country and the amount of tracking flows confined within its borders, but there are other important reasons that determine the level of national confinement, as we explain in Sect. 5. In Sect. 7, we use large ISP datasets to cover additional central and north European countries for which we have rather few users in the Sankey diagram to further investigate if the number of users can influence the confinement level for such countries.

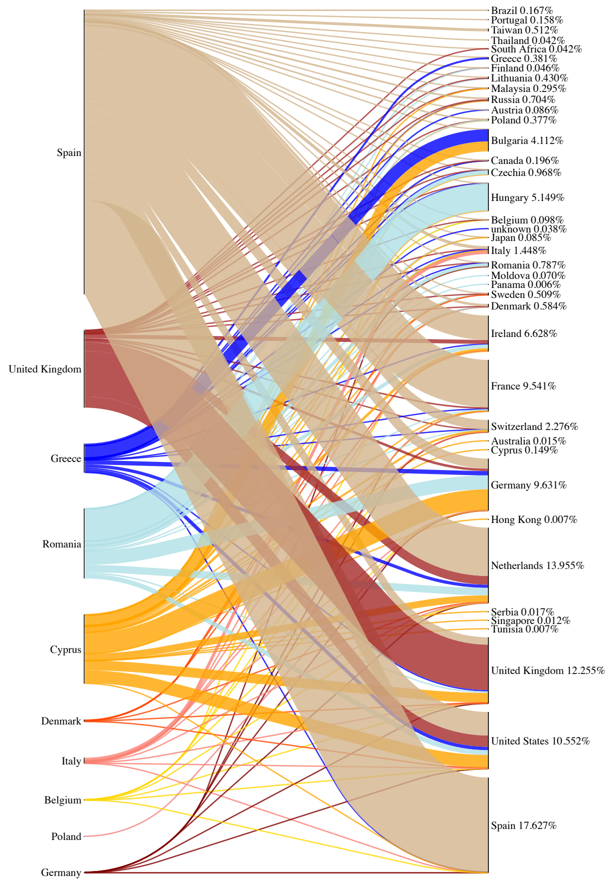


Figure 8: The flow of ad and tracking domains from European Union (EU28) countries using the RIPE IPmap geolocation service.

5 KEEPING TRACKING FLOWS LOCAL

In this section, we look at the effectiveness of different methods for improving the localization of tracking flows. We consider two methods to increase localization, namely, (i) DNS redirection, and (ii) mirroring of tracking PoPs.

Apart from its value for privacy, localization can be beneficial also for the ad domains, especially those serving targeted ads using

Table 5: Potential localization improvements under different scenarios.

EU28 - 1,824,873	Percentage In		Improvement	
	Country	Cont.	Country	Cont.
Default	27.60%	88.00%	-	-
Redirections (FQDN)	52.15%	93.53%	24.55%	5.53%
Redirections (TLD)	66.13%	98.33%	38.53%	10.33%
POP Mirroring (Cloud)	30.79%	92.09%	3.19%	4.09%
Redirection (TLD) + POP Mirroring (Cloud)	68.12%	99.20%	40.52%	11.20%

Table 6: Potential localization improvement over TLD optimizations for EU28 countries using alternative large public Cloud PoPs.

PoP Mirroring (Cloud) Over Redirection (TLD)			Migration to Cloud Over Redirection (TLD)	
Country	# Requests	% Impr.	Country	% Impr.
UK	261,915	5.47	Denmark	96.85
Spain	961,231	1.84	Greece	79.25
Greece	98,281	1.29	Romania	72.12
Italy	19,801	1.14	Italy	25.64
Romania	236,528	1.13	UK	18.20
Cyprus	234,433	0	Spain	12.15
Denmark	7,503	0	Cyprus	0

the RTB protocol. In RTB delivery delays need to be kept low to improve the performance of real time bidding.

5.1 Localization potential using DNS

Our first investigation involves a simple DNS redirection based on alternative servers that we have observed in our dataset for the same tracking domain. We first quantify the improvement potential by looking for alternative server locations operate under the same fully qualified domain names (FQDN). Then, we find the corresponding TLD for each FQDNs and consider the case of redirecting requests for the FQDN to any alternative servers that belong to the same TLD level that can further improve the confinement.

Table 5 depicts the results of the different approaches. The first row (Default) depicts the base line of the confinement percentage at country and continent level for all the tracking flows that we observe in our dataset. In the case of DNS redirections based on FQDN level, we observe an additional confinement up to 5.5% and 24.55% at continent and country level, respectively (Table 5 - Row 2, Right column). DNS Redirection has a non-negligible positive contribution to keeping tracking local within GDPR jurisdiction. If applied at TLD, the improvement in our dataset is more than 10%. However, it plays an even higher role in improving confinement within national boarders. In this case, the improvement under TLD redirection is an impressive 38%.

Based on our "what-if" analysis, we conclude that, with a more thoughtful (or GDPR friendly) DNS redirection on behalf of the tracking domain administrators, the overall confinement can be improved at both country and continent level, with minimal additional financial cost (that includes additional server and network capacity as appropriate). Note that with DNS redirection, it is easy to change the assignment of users to a server IP. For example, google time to live (TTL) for DNS records is 300 seconds and facebook TTL is 7,200 seconds. Thus, DNS redirection can take place in relatively small time scale, from seconds to a few hours.

5.2 Localization potential using Mirroring

For our second investigation, we turn our attention to PoP mirroring using cloud services and the potential localization that such an optimization can offer. For this hypothetical setup, we collect information from nine major cloud service providers in which we know from our dataset that tracking domains lease servers. These public clouds make their global footprint and, in some cases, the associated IP ranges publicly available in order to: (i) attract new customers by advertising their presence at different regions, (ii) improve the operation of current customers by providing an accurate and up-to-date map of IP ranges to physical location, and (iii) to white-list the IP ranges, *e.g.,* to update firewall rules. The major cloud providers we consider in this study are: Amazon AWS [2], Microsoft Azure [17], IBM Cloud [24], CloudFlare [23], Digital Ocean [6], Equinix [8], Oracle Cloud [18], Rackspace [19], and Google Cloud [11]. For each cloud service we collect the physical location of their operational datacenters, at a country level, as advertised in each cloud service website.

First, we check if the the confinement within the user's region can be further improved if tracking domains that are already hosting their server on these cloud services utilize additional PoPs (PoP Mirroring), *i.e.,* different datacenters of the same cloud service provider. Under the "PoP Mirroring" scenario (Table 5 - Row 3), it is evident that PoP mirroring yields good improvement of confinement within the GDPR legislation region, but not so great on the national level. Furthermore, we observe that many countries lack large public cloud PoPs, and the improvement in confinement is expected to be marginal for these countries. Finally, at Table 5 - Row 5, we present the confinement percentage and improvement, respectively, by combining DNS redirection at TLD level with PoP Mirroring. The combination yield an additional improvement of 40.52% and 11.2% at the country and continent level, respectively.

Next, we investigate the extreme scenario where all tracking domains can potentially migrate to any cloud PoP from all PoPs that we observe in all nine major cloud services. After examining our results in Table 6 (Right column), we see that countries such as, Denmark (69.85%), Greece (79.25%) and Romania (72.12%) can achieve 96.85%, 79.25% and 72.12% additional confinement over the "Default" outgoing tracking flows, respectively. In contrast, using only PoP Mirroring (Table 6 - Right column) the confinement improvement is negligible, below 1.3%, for the above three countries. On the other hand, countries such us Cyprus cannot benefit from this scenario since none of the nine cloud services in our study has a presence in the country. Note that if a tracking operator is willing to utilize any datacenter available in a country, then it is possible to achieve complete flow confinement at the national level. In all EU28 countries there is at least one datacenter, even in the smallest country.

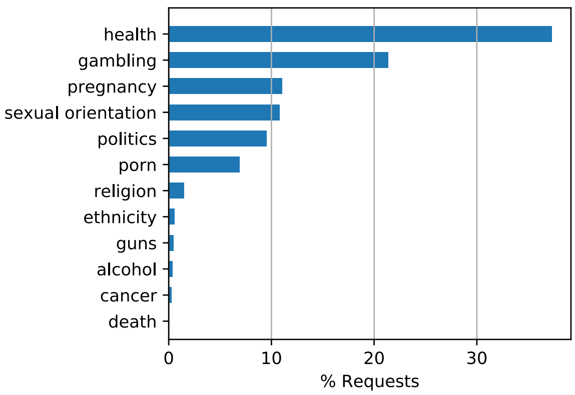


Figure 9: The percentage of websites for each sensitive topic that ad + tracking third-party domains where present in our dataset. We observe 127K requests towards sensitive topics, 2.89% of the total tracking flows we observed

In summary, we observe that there exists a correlation between the density level of IT infrastructure of a country, mostly in terms of datacenters, and the confinement of tracking flows within its borders. The confinement of tracking flows within national borders can be improved in many cases, either by using DNS or mirroring of tracking PoPs, at a relatively low cost. However, in some small countries with less developed IT infrastructure, the improvement of the confinement of tracking flows within national borders may require proportionally high cost or expansion of the footprint of major cloud providers in these countries.

6 TRACING SENSITIVE TRACKING FLOWS

GDPR [5] defines sensitive personal data as any data *"revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership*", also *"genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation"*. In this section, we try to find if tracking flows exist on sensitive data, and if they do, look at their geographic confinement.

6.1 Methodology

In total, we observe more than 76K first party domains in our dataset. To identify domains that fall into the sensitive categories we use a multi-stage filtering process involving automated and manual inspection of website content.

As a first step we use AdWords [10], an online tagging service provided by Google, to detect the interest topics of the visited domain. Usually we have 5 to 15 interest topics per domain. Next, we use automated look up to detect whether any of the AdWords categories of a specific domain contains any of the 7 sensitive categories defined by GDPR. If a domain topic matched we include it in our analysis. We also manually examine the remaining domains to see if they contained any semantic categories that had a semantic relevance/overlap with GDPR defined sensitive terms. We used multiple people for this and include a domain in our analysis when at least 2 independent examiners agreed that it was relevant to a GDPR sensitive term.

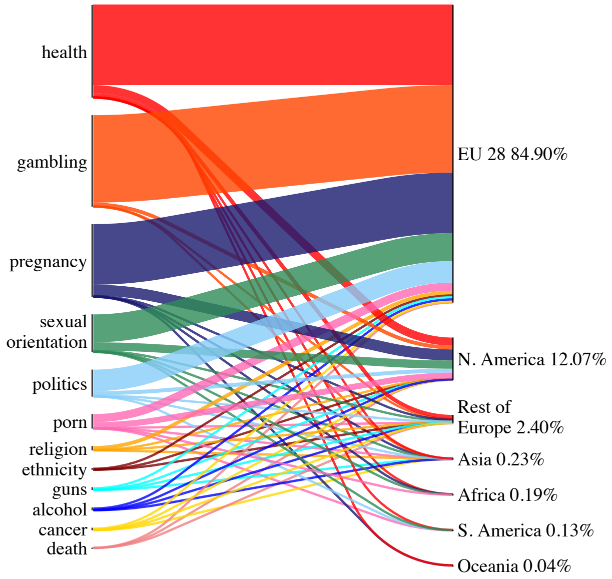


Figure 10: The destination continent of tracking flows for each sensitive category using EU28 users.

Overall, we inspected 5,698 domains over a period that spans two weeks. We chose to manually inspect the content since most tagging systems do not include sensitive categories. For example a website related to pregnancy falls into the category "Health". Similarly, websites related to pornography, alcohol and gambling will fall into the categories "Men's Interests", "Food & Drinks" and "Games", respectively. Thus, by manually inspecting the website content we can identify websites belonging to sensitive categories with high accuracy. In total we identify 12 sensitive categories (see Fig. 9) from 1,067 domains. The total number of tracking flows related to sensitive categories is 127K.

6.2 Results

Figure 9 depicts the percentage of tracking flows for each sensitive category. The most heavily tracked category is "Health" with 38% of the tracking flows followed by gambling with 22%. Sex related categories, such as, sexual orientation and pregnancy have identical percentage ≈ 11%, followed by politics and porn at 9% and 7%, respectively. Religion, ethnicity, guns, alcohol, cancer and death are below 3%. Note that in the case of the categories cancer and death, both belong to the category "Health", but we report them separately due to their obvious sensitivity.

In Fig. 10, we present the destination continent of the tracking flows for each sensitive category. We observe similar trends as with the aggregated results, *i.e.,* most tracking flows are confined within GDPR (EU28 84.9%) but a non-trivial percentage (12.7%) is collected

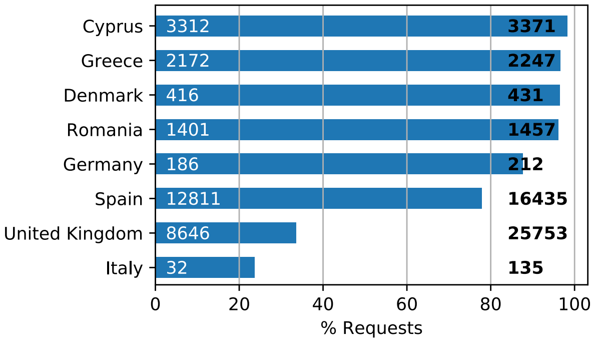


Figure 11: The percentage of tracking flows from sensitive websites that travels outside the users' country using users within EU28 countries.

in North America. The categories with the highest leakage out of EU28 are: porn (44%), sexual orientation (36%) and alcohol (33%).

Finally, in Fig. 11, we plot the confinement for each EU28 country, where we observe tracking flows on sensitive category domains. The black numbers (right) depict the total number of sensitive flows for the corresponding country, and the white numbers (left) show the flows that travel outside the country. The trends are similar to the aggregated results thus, countries with a small population and limited IT infrastructure, *e.g.,* Cyprus, Greece, Denmark and Romania seem to experiencing more leakage on sensitive tracking flows.

7 SCALING UP: A VIEW FROM ISPs

Next, we examine the geographical distribution of tracking flows involving subscribers of various large European ISPs. In particular, we analyze data from four ISPs in three European countries. The analysis of ISP data contributes to our study in multiple ways:
(i) the ISP datasets capture the traffic of millions of real users, thereby allowing us to *scale up our study* and validate our previous observations and conclusions drawn from our browser extension users, (ii) they *increase the diversity of our study*, not only because the studied ISPs operate in different countries, but also because their users are residential, mobile, or both, and (iii) they operate in countries where we did not have a large user base in our active experiment, thus, *complementing our study.*

7.1 Profile of ISPs

Table 7 provides a brief summary of the profile of the four ISPs.

DE-Broadband: This is one of the largest ISPs, in terms of both customer base and traffic volume in Germany and Europe with more than 15 million broadband residential lines. Since it is difficult to estimate the number of users that take Internet access from these lines, we refer to the number of broadband households.

DE-Mobile: This is one of the largest mobile providers, in terms of both customer base and traffic volume, in Germany and in Europe with more than 40 million subscribers.

Table 7: Profile of the four European ISPs in our study.

Name	Country	Demographics
DE-Broadband	Germany	15+ million broadband households
DE-Mobile	Germany	40+ million mobile users
PL	Poland	11+ million mobile and broadband users
HU	Hungary	6+ million mobile and broadband users

PL: This is one of the largest mobile and broadband ISPs in Poland, both in terms of customers and traffic volume, that offers both mobile and broadband services. Overall, it has more than 11 million mobile and broadband users.

HU: This is one of the largest mobile providers in Hungary, that has also a smaller fraction in the broadband market. Overall, this ISP serves more than 6 million users in Hungary, primarily mobile users.

7.2 Methodology

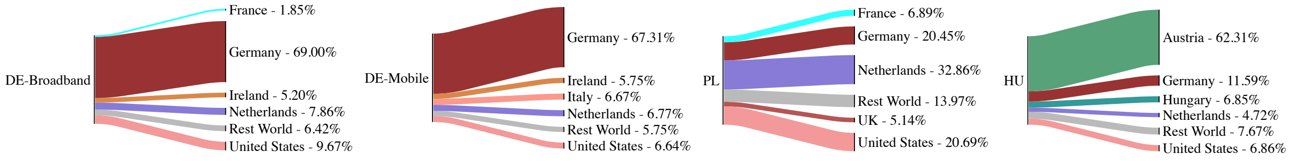
To identify the tracking flows from ISP NetFlows, we rely on the list of IPs of tracking services compiled using the browser extension as described in Sect. 3. In addition, we also collected data for the period mid-Jan. to July 2018 using the same methodology. We perform the ISP study using daily snapshot activity, on four days: (i) Wednesday, Nov. 11, 2017, (ii) Wednesday, April 4, 2018, (iii) Wednesday, May 16, 2018 (close to the implementation date of the EU GDPR law on May 25, 2018), and (iv) Wednesday, June 20, 2018 (after the implementation date of the EU GDPR law). Note that the data collection in the time period between mid- Jan. to end of July is related only to the results presented in table 7 under the columns June 20 and is only related to the ISP's analysis after the implementation date of GDPR. The data collected in the above time period are not included in the data analysis in Sect. 4.

Our daily snapshots consist of 24 hour NetFlow [32] data collected at both network edges, internal (*e.g.,* end-users) as well as external (*i.e.,* peering links). The NetFlow data provides per flow the collection timestamp, exporting router and interface identifiers, the layer-4 transport protocol, the source and destination IPs and protocol ports, the IP type of service field as well as sampled number of packets and bytes. The NetfFlow sampling rate is constant throughout the experiment. For our study, we consider only the router interfaces that carry user traffic, *i.e.,* internal network edge routers. All the ISPs perform ingress network filtering (BCP38 and RFC2827 [33]) against spoofing. We noticed that the majority of the flows (more than 99.5%) that involve tracking IPs are Web traffic in ports 80 or 443, using either TCP or UDP (due to the increasing usage of QUIC [40, 54]) protocols. Overall, more than 83% of the traffic used port 443, thus, it was encrypted.

Ethical considerations: To protect the privacy of users, the IPs of the end users in the Netflow data are anonymized, *i.e.,* replaced with the country code where each ISP operates. We do not collect, store, or process any information regarding the users. For our study, individual user IPs and activity are not important considering we know that the users are located in the country that the ISP operates. To report on the number of flows that involve the tracking IPs, we use a hash function to check if the source or the destination of

Table 8: Sampled tracking flow statistics across EU ISPs and over time.

	DE-Broadband				DE-Mobile				PL				HU			
	Nov 8	April 4	May 16	June 20	Nov 8	April 4	May 16	June 20	Nov 8	April 4	May 16	June 20	Nov 8	April 4	May 16	June 20
#Sampled Tracking Flows (in Millions)	1,057.0	1,200.8	1,105.3	963.4	70.4	77.4	70.8	74.5	13.8	13.8	12.4	11.9	43.3	50.2	39.3	33.6
EU28	**88.5%**	**87.7%**	**86.5%**	**88.3%**	**91.1%**	**90.8%**	**89.9%**	**92.5%**	**77.5%**	**75.6%**	**74.7%**	**75%**	**89.5%**	**93.1%**	**92.4%**	**91.6%**
N. America	10%	9.3%	9.2%	8.4%	6.9%	6.6%	6.4%	5.1%	19.8%	21.5%	22%	21.3%	10.2%	6.3%	7%	7.7%
Rest Europe	<1%	1.7%	2.9%	1.8%	<1%	2%	3.1%	1.3%	1.9%	1.9%	1.7%	3.4%	<1%	<1%	<1%	<1%
Asia	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%
Rest World	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	1.1%	<1%	<1%	<1%	<1%	<1%



(a) DE-Broadband (Germany)

(b) DE-Mobile (Germany)

(c) PL (Poland)

(d) HU (Hungary)

Figure 12: The top 5 countries for each ISP dataset where the tracking flows are terminating (April 4).

the flow matches any tracking IP. If it matches, we increase the counter for this tracking IP by one. For our analysis we follow the methodology described in Sect. 4 to infer border crossing.

7.3 Results

We now turn our attention to the assessment of the confinement of tracking flows within EU GDPR (EU28) and national borders. For a summary of results, we refer to Table 8. Notice that the sampled tracking flows are in the order of multiple millions, but the estimated number of tracking flows is several orders of magnitude larger. For example, the estimated number of tracking flows for DE-Broadband on April 4, 2018 is more than 1 Trillion flows. This highlights the large number of flows that are dedicated to tracking, which accounts for, in the case of DE-Broadband, around 3% of the total flows in this ISP. It is also worth mentioning that the number of tracking flows in mobile operators, *e.g.,* DE-Mobile, is relatively lower. This happens because Web activity in mobile is lower than in fixed, since much of the traffic goes over smartphone apps instead of browsers.

Baseline results: Overall, the analysis of the four large European ISPs shows comparable confinement ratios as those reported based on browser extension data. Indeed, the analysis of tracking flows observed by 183 users in EU28 countries over a period of four months (see Sect. 4) and the post GDPR period between mid-Jan.-July 2018 showed that around 85% of the tracking flows terminated within EU28 borders. As shown in Table 8, the confinement of tracking flows within EU28 as observed from more than 60 million European users in three EU28 countries for the same period ranges from 76% to 93%, which is in pretty good agreement with the results of Fig. 7(b) derived based on browser extension data. When focusing on the difference across time, we observe that the confinement of tracking flows within EU28 has not changed dramatically in the last six months, and it has been high throughout this period as well as before the EU GDPR implementation date (May 25, 2018). Similar observations apply for June 20, 2018 after the EU GDPR implementation date. This is an indication that many companies in the ad and tracking space took measures to confine tracking flows within EU28 borders according to GDPR law.

The effect of provider type: When comparing the confinement across networks, there are some noticeable differences. The ISPs that are primarily mobile operators, namely DE-Mobile and HU, yield higher confinement (above 90%). This is to be expected as mobile users typically rely on the DNS service of their provider, and, thus get mapped to nearby tracking servers more frequently, if available. On the other hand, broadband users increasingly rely on third-party DNS services [46], *e.g.,* Google DNS, Quad9, Level3, *etc.,* and thus, may be mapped to available servers in different countries and regions.

The effect of local IT infrastructure: We also assess the extend of which local IT infrastructure deployment plays an important role in increasing the confinement of tracking flows within national borders. In Fig. 12, we show the confinement of tracking flows in the top five countries for the four ISPs on April 4, 2018; similar observations are derived for the other two dates in our dataset. The two ISPs that operate in Germany, a country with very developed IT and networking infrastructure in Europe, have considerably higher confinement within national borders, 69% for DE-Broadband and 67.31% for DE-Mobile, compared to 0.25% (not visible in Fig. 12(c)) and 6.85% for PL and HU, respectively. As expected, a large fraction of tracking flows that cross borders are served by servers in other neighboring EU countries, with a heavy bias on countries with advanced IT infrastructure, such as the Netherlands and Ireland in the case of German operators, Germany and the Netherlands in the case of PL, Austria in the case of HU. This analysis agrees with the analysis of the data collected from the real users using the browser extension (see Sect. 4).

8 RELATED WORK

A subsequential amount of recent work has studied the privacy implications of online advertising and web tracking in desktop [28–30, 35, 41, 51, 58, 61], mobile [53, 60] or mixed platforms [36, 42, 52].

Table 9: The comparison table of the related work and their corresponding key features

		[52]	[36]	[29]	[58]	[30]	[42]	[53]	[41]	[35]	[61]	[28]	[60]	[51]	**This Work**
★ Request Classification	ABP	●	●	●	●	●	●			●	●			●	●
	Ghostery		●												
	Custom list						●	●	●				●		
	Other	✓ Custom Corrections			✓ Custom Corrections				† Cookies based			† Text Ads			✓ Custom Corrections
Requests Type	Ads	●	●	●	●	●				●	●	●	●	●	●
	Tracking	●	●		●	●	●	●	●	●				●	●
Measurement Type	Active	●	●	●	●	●	●	●	●	●	●	●		●	●
	Passive								●				●	●	●
Platform Type	Desktop	●	●	●	●	●	●		●	●	●	●		●	●
	Mobile	●	† User agent				●	●					●		●
Data Collection	Crawling		†	†	†	†		†	†	†	†	†		†	
	Real Users	✓													✓
	Other						† Control environment	● Apps Store					● Net Traces		● Net Flows
Infrastructure Geolocation	MaxMind	†	†	†											
	Other		† WHOIS Legal Entities					† Legal Entities							✓ RIPE IPmap
Traffic Type	HTTPS	✓	✓	✓	✓	✓			✓	✓	✓	✓			✓

✓ Positive, † Negative, ● Neutral. ★ For AdBlockPlus (ABP) and Ghostery filter lists, additional corrections are required depending on the use case, noted as "✓ Custom Corrections".

In Table 9, we summarize and compare some of the key features and approaches from this literature. We highlight all positive features of each work with a green checkmark, all negative ones with a red cross, and all neutral ones with a black dot. The rating scheme in use is based on the challenges stated in Sect. 2.3. To the best of our knowledge, our work is the only existing work fully dedicated to the study of cross border tracking on the web. Few other works have touched upon geographic matters of tracking but never at the depth and breadth that we have. Also, in order to carry our study, we had to come up with methodological contributions that progress the aggregate state of the art from previous work along multiple directions (see the right column of Table 9 for an overview).

In regards to reviewing the literature, we focus on four key aspects:

Third-party request classification: One of the most critical and non-trivial problems is to be able to distinguish a third-party request either as ad- or tracking-related or not. Currently, the most common solution is to use the *"easylist"* and *"easyprivacy"* [7] lists to detect ad- and tracking-related requests, respectively. A naive usage of the above lists can lead to an over- or under-estimation of the third-party requests belonging to each category depending on how the lists are used. For example, one can consider all the domains included in the list. This approach will lead to an overestimation since domains such as *"google.com"* can serve all three types of request. Another way of using the lists is by using the included blocking rules and classify the third-party request only when there is an exact match. Note that the lists are constructed to block third-party requests as observed from the real users browser, thus, any subsequent third-party requests that is initiated by the blocked content may include additional domains that will stay outside the list rules as explained earlier in Sect. 3.2. The above observations are also identified and reported in [52, 58] and are also confirmed by our own work. We refer to the extra work we do to collect additional trackers that do not appear in the standard lists as "Custom Corrections" in Table 9.

Data collection: The data collection process (Table 9 - Row 5) can also influence the results in some cases. The convenience of using web crawling as oppose to real users can limit the number of observable third-party requests due to the lack of user interaction (scroll or page down) on a webpage that includes tracking code. To improve user experience, reduce data consumption, and charge advertisers accurately, ads are rendered only when the ad slot becomes visible to the user. For more discussion about the advantages of using real users in the mobile environment see [42, 52]. Web crawling is a better approach (given that it is faster) for studies that do not require user interaction with the content, *e.g.,* collecting information about mobile apps from app stores [53] or from web-archives [41]. In this work, the data collection takes place on real users' browsers using a browser extension to overcome the above limitation as described in Sect. 3.1.

Infrastructure geolocation: In order to improve the user experience, most web platforms and e-stores use a geolocation service for things like customizing content language or currency based on the location of a visitor. As a result, most geolocation services turn their attention towards the accurate geolocation of end users connecting from residential or mobile broadband networks. Accurately geolocating server infrastructure is a secondary priority for such services. Indeed, by manually examining some of the available geolocation services, we noticed that the location for most of the IPs related to infrastructure servers was determined based on the legal entity owner's location (see Sect 3.4). Thus, using such services to infer server location is problematic. If the focus of a study is to only geolocate the legal entity behind a specific server IP, then these services can be used safely (Table 9 - Row 6) [36, 53]. In this work we identify the above problem and we avoid it by utilizing a state-of-the-art solution based on active measurement to correctly geolocate infrastructure servers involved in web tracking and advertising activities.

Traffic type: An additional advantage is to have a methodology that can work on encrypted traffic (Table 9 - Row 7). Most ad and tracking related third-party requests that we observe in our study have already moved to encrypted traffic (83.14% based on the real users dataset). As we can see in Table 9, ten out of fourteen studies are able to operate on encrypted traffic. In this work we propose a novel methodology that can identify tracking flows in the wild using ISPs NetFlows. In more details, we use active measurement to carefully identify the IPs associated with ad and tracking related activities within real users browsers (see Sect. 3.3) irrespectively of the protocol used (HTTP or HTTPS) and use this information to analyze ISPs NetFlows at the IP level (see Sect. 7.2) avoiding the need of any additional meta-data or contextual information.

9 CONCLUSION

We have developed an elaborate measurement and analysis methodology for quantifying the percentage of tracking flows that terminate within national borders. Our analysis reveals that most tracking flows on European Union citizens terminate within EU members thus putting them under the full jurisdiction of GDPR and permitting European data protection authorities to conduct full investigations. This is a rather optimistic result when contrasted with what happens in other continents, *e.g.,* South America, that has most of the tracking flows on its citizens terminating in North America.

Naturally, the level of confinement within national borders is substantially lower than continent-wide confinement. Looking at individual European countries we see a clear correlation between the size of a country and the amount of tracking that is confined within its borders. There also exists a correlation between the density level of IT infrastructure of a country, mostly in terms of datacenters, and the confinement of tracking flows within its borders. National confinement can be improved substantially via simple DNS redirection to alternative tracking end points. This is something that most tracking companies could implement with a rather small cost. However, for some smaller countries with less advanced IT infrastructure, DNS redirection alone is not enough but rather needs to be paired with tracking PoP mirroring within the country.

An important finding of our study is that the confinement level of tracking flows relating to protected data categories is similar to that of general traffic. This is a positive or negative result depending on one's view-point: positive in the sense that such tracking can readily be investigated since most of it terminates within GDPR jurisdiction; negative in the sense that some of it should not be occurring in the first place.

In this work we provide a methodology on how to bootstrap and scale an experiment to detect and geolocate the different ad and tracking related stakeholders under the new regulations with regards to data protection and we apply it to the timely topic of GDPR. We can continuously monitor the compliance to GDPR over time and also include the monitoring of other regulations in the future at different regional (*e.g.,* USA) or content scope (Children's Online Privacy Protection Act - COPPA [3, 38], *etc.*)

In our future work we intend to build a system around our methodology, deploy it, and make it available to whomever would like to have hard data on cross-border tracking in real-time and at scale. We also plan to extend our methodology to go beyond the terminating end-point of tracking to capture inter-tracker collaboration and data exchange.

ACKNOWLEDGMENTS

We thank the anonymous reviewers and our shepherd Aruna Balasubramanian for their thoughtful reviews and suggestions. We also thank Nicolas Kourtellis (Telefonica I+D) for his help during the users recruitment process on the CrowdFlower platform. Finally, we would like to thank Massimo Candela and Emile Aben (RIPE NCC) for their support and help in using the RIPE IPmap geolocation tool.

This work and its dissemination efforts were supported in part by the European Research Council (ERC) grant ResolutioNet (ERC-StG-679158) and by the European Union H2020 project SMOOTH (Grant agreement No.:786741).

REFERENCES

[1] AdBlock Plus - Surf the web without annoying ads! https://adblockplus.org/.
[2] Amazon - AWS IP Address Ranges in JSON format. https://docs.aws.amazon.com/general/latest/gr/aws-ip-ranges.html.
[3] Children's Online Privacy Protection Act (COPPA). https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule.
[4] CrowdFlower. https://www.crowdflower.com/.
[5] Data protection in the EU, The General Data Protection Regulation (GDPR); Regulation (EU) 2016/679. https://ec.europa.eu/info/law/law-topic/data-protection/.
[6] Digital Ocean network. https://status.digitalocean.com/.
[7] Easylist - The primary filter list that removes most adverts from international webpages. https://easylist.to/.
[8] Equinix: Global Data Centers and Colocation Services. https://www.equinix.com/locations/.
[9] Ghostery - Makes the Web Cleaner, Faster and Safer! https://www.ghostery.com/.
[10] Google AdWords. https://adwords.google.com/.
[11] Google Cloud Locations. https://cloud.google.com/about/locations/.
[12] Google: Our Infrastructure. https://peering.google.com/#/infrastructure.
[13] Google: Real-Time Bidding Protocol. https://developers.google.com/ad-exchange/rtb/start.
[14] Interactive Advertising Bureau: OpenRTB (Real-Time Bidding). https://www.iab.com/guidelines/real-time-bidding-rtb-project/.
[15] IP-API - Free Geolocation API. http://ip-api.com/.
[16] MaxMind: IP Geolocation and Online Fraud Prevention. https://www.maxmind.com.
[17] Microsoft Azure Datacenter IP Ranges. https://www.microsoft.com/en-us/download/details.aspx?id=41653.
[18] ORACLE: Data Regions for Platform and Infrastructure Services. https://cloud.oracle.com/data-regions.
[19] Rackspace Global Infrastructure. https://www.rackspace.com/about/datacenters.
[20] RIPE Atlas. https://atlas.ripe.net/.
[21] RIPE NCC OpenIPmap: Geolocating Internet Infrastructure with Inference Engines and Crowdsourcing. https://ipmap.ripe.net/.
[22] Robtex - Everything you need to know about domains, DNS, IP, Routes, Autonomous Systems, and much, much more! https://www.robtex.com/.
[23] The Cloudflare Global Anycast Network. https://www.cloudflare.com/network/.
[24] The IBM Cloud network. https://www.ibm.com/cloud-computing/bluemix/our-network.
[25] OpenRTB API Specification Version 2.3.1. https://www.iab.com/wp-content/uploads/2015/05/OpenRTB_API_Specification_Version_2_3_1.pdf, 2015.
[26] Internet Advertising Bureau: Advertising Revenue Report. https://www.iab.com/insights/iab-internet-advertising-revenue-report, 2018.
[27] G. Acar, M. Juarez, N. Nikiforakis, C. Diaz, S. Gürses, F. Piessens, and B. Preneel. FPDetective: Dusting the Web for Fingerprinters. In *ACM CCS*, 2013.
[28] R. Balebako, P. L. G. De León, R. Shay, B. Ur, Y. Wang, and L. F. Cranor. Measuring the Effectiveness of Privacy Tools for Limiting Behavioral Advertising. In *W2SP Workshop*, 2012.
[29] P. Bangera and S. Gorinsky. Ads versus Regular Contents: Dissecting the Web Hosting Ecosystem. In *IFIP Networking*, 2017.
[30] M. A. Bashir, S. Arshad, E. Kirda, W. Robertson, and C. Wilson. How Tracking Companies Circumvent Ad Blockers Using WebSockets. In *Workshop on Technology and Consumer Protection*, 2018.
[31] M. Calder, X. Fan, Z. Hu, E. Katz-Bassett, J. Heidemann, and R. Govindan. Mapping the Expansion of Google's Serving Infrastructure. In *ACM IMC*, 2013.
[32] B. Claise. Cisco Systems NetFlow Services Export Version 9, October 2004. IETF RFC 3954.
[33] P. Ferguson and D. Senie. Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing, May 2000. IETF RFC 2827.

[34] M. J. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan. Geographic Locality of IP Prefixes. In *ACM IMC*, 2005.

[35] N. Fruchter, H. Miao, S. Stevenson, and R. Balebako. Variations in Tracking in Relation to Geographic Location. *CoRR*, 2015.

[36] A. Gervais, A. Filios, V. Lenders, and S. Capkun. Quantifying Web Adblocker Privacy. 2017.

[37] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos. A Look at Router Geolocation in Public and Commercial Databases. In *ACM IMC*, 2017.

[38] I. Reyes and P. Wijesekera and A. Razaghpanah and J. Reardon, N. Vallina-Rodriguez and S. Egelman and C. Kreibich. Is Our Children's Apps Learning? Automatically Detecting COPPA Violations. In *Workshop on Technology and Consumer Protection (ConPro)*, 2017.

[39] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe. Towards IP geolocation using delay and topology measurements. In *ACM IMC*, 2006.

[40] A. Langley, A. Riddoch, A. Wilk, A. Vicente, C. Krasic, D. Zhang, F. Yang, F. Kouranov, I. Swett, J. Iyengar, J. Bailey, J. Dorfman, J. Roskind, J. Kulik, P. Westin, R. Tenneti, R. Shade, R. Hamilton, V. Vasiliev, W-T. Chang, and Z. Shi. The QUIC Transport Protocol: Design and Internet-Scale Deployment. In *ACM SIGCOMM*, 2017.

[41] A. Lerner, A. Kornfeld Simpson, T. Kohno, and F. Roesner. Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016. In *USENIX Security Symposium*, 2016.

[42] C. Leung, J. Ren, D. Choffnes, and C. Wilson. Should You Use the App for That?: Comparing the Privacy Implications of App- and Web-based Online Services. In *ACM IMC*, 2016.

[43] M. Falahrastegar and H. Haddadi and S. Uhlig and R. Mortier. The Rise of Panopticons: Examining Region-Specific Third-Party Web Tracking. In *TMA*, 2014.

[44] M. Falahrastegar and H. Haddadi and S. Uhlig and R. Mortier. Tracking Personal Identifiers Across the Web. In *PAM*, 2016.

[45] J. R. Mayer and J. C. Mitchell. Third-party Web Tracking: Policy and Technology. In *IEEE Symposium on Security and Privacy*, 2012.

[46] J. S. Otto, M. A. Sanchez, J. P. Rula, and F. E. Bustamante. Content delivery and the natural evolution of DNS - Remote DNS Trends, Performance Issues and Alternative Solutions. In *ACM IMC*, 2012.

[47] P. Papadopoulos, N. Kourtellis, and E. P. Markatos. Exclusive: How the (synced) Cookie Monster breached my encrypted VPN session. In *European Workshop on Systems Security*, 2018.

[48] P. Papadopoulos, P. Rodriguez, N. Kourtellis, and N. Laoutaris. If you are not paying for it, you are the product: how much do advertisers pay to reach you? In *ACM IMC*, 2017.

[49] J. Parra-Arnau, J. P. Achara, and C. Castelluccia. *MyAdChoices*: Bringing Transparency and Control to Online Advertising. *TWEB*, 2017.

[50] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye. IP Geolocation Databases: Unreliable? *ACM CCR*, 41(2), 2011.

[51] E. Pujol, O. Hohlfeld, and A. Feldmann. Annoyed Users: Ads and Ad-Block Usage in the Wild. In *ACM IMC*, 2015.

[52] A. Razaghpanah, R. Nithyanand, N. Vallina-Rodriguez, S. Sundaresan, M. Allman, C. Kreibich, and P. Gill. Apps, Trackers, Privacy, and Regulators: A Global Study of the Mobile Tracking Ecosystem. In *NDSS*, 2018.

[53] B. Reuben, L. Ulrik, M. Van Kleek, J. Zhao, T. Libert, and N. Shadbolt. Third Party Tracking in the Mobile Ecosystem. *CoRR*, 2018.

[54] J. Ruth, I. Poese, C. Dietzel, and O. Hohlfeld. A First Look at QUIC in the Wild. In *PAM*, 2018.

[55] InMon – sFlow. http://sflow.org/.

[56] S. S. Siwpersad, B. Gueye, and S. Uhlig. Assessing the geographic resolution of exhaustive tabulation for geolocating Internet hosts. In *PAM*, 2008.

[57] O. Starov, P. Gill, and N. Nikiforakis. Are You Sure You Want to Contact Us? Quantifying the Leakage of PII via Website Contact Forms. *PoPETs*, 2016.

[58] E. Steven and A. Narayanan. Online Tracking: A 1-million-site Measurement and Analysis. In *ACM CCS*, 2016.

[59] F. Streibelt, J. Boettger, N. Chatzis, G. Smaragdakis, and A. Feldmann. Exploring EDNS-Client-Subnet Adopters in your Free Time. In *ACM IMC*, 2013.

[60] N. Vallina-Rodriguez, J. Shah, A. Finamore, Y. Grunenberger, K. Papagiannaki, H. Haddadi, and J. Crowcroft. Breaking for Commercials: Characterizing Mobile Advertising. In *ACM IMC*, 2012.

[61] R. J. Walls, E. D. Kilmer, N. Lageman, and P. D. McDaniel. Measuring the Impact and Perception of Acceptable Advertisements. In *ACM IMC*, 2015.

[62] J. Wang, W. Zhang, and S. Yuan. Display Advertising with Real-Time Bidding (RTB) and Behavioural Targeting. *Foundations and Trends in Information Retrieval*, 11, Oct 2016.

[63] F. Weimer. Passive DNS Replication. In *17th Annual FIRST Conference*, 2005.