Clash of the Trackers: Measuring the Evolution of the Online Tracking Ecosystem

Preprint · July 2019

4 authors:

Konstantinos Solomos
University of Illinois at Chicago

Panagiotis Ilia
Foundation for Research and Technology - Hellas

Sotiris Ioannidis
Foundation for Research and Technology - Hellas

Nicolas Kourtellis
Telefónica I+D

Clash of the Trackers: Measuring the Evolution of the Online Tracking Ecosystem

Konstantinos Solomos
FORTH,Greece
solomos@ics.forth.gr

Panagiotis Ilia
Univ. of Illinois at Chicago, USA
pilia@uic.edu

Sotiris Ioannidis
FORTH,Greece
sotiris@ics.forth.gr

Nicolas Kourtellis
Telefonica Research, Spain
nicolas.kourtellis@telefonica.com

ABSTRACT

Websites are constantly adapting the methods used, and intensity with which they track online visitors. However, the wide-range enforcement of GDPR since one year ago (May 2018) forced websites serving EU-based online visitors to eliminate or at least reduce such tracking activity, given they receive proper user consent. Therefore, it is important to record and analyze the evolution of this tracking activity and assess the overall "privacy health" of the Web ecosystem and if it is better after GDPR enforcement. This work makes a significant step towards this direction. In this paper, we analyze the online ecosystem of 3rd-parties embedded in top websites which amass the majority of online tracking through 6 time snapshots taken every few months apart, in the duration of the last 2 years. We perform this analysis in three ways: 1) by looking into the network activity that 3rd-parties impose on each publisher hosting them, 2) by constructing a bipartite graph of "publisher-to-tracker", connecting 3rd parties with their publishers, 3) by constructing a "tracker-to-tracker" graph connecting 3rd-parties who are commonly found in publishers. We record significant changes through time in number of trackers, traffic induced in publishers (incoming vs. outgoing), embeddedness of trackers in publishers, popularity and mixture of trackers across publishers. We also report how such measures compare with the ranking of publishers based on Alexa. On the last level of our analysis, we dig deeper and look into the connectivity of trackers with each other and how this relates to potential cookie synchronization activity.

1 INTRODUCTION

Online users' privacy is constantly violated by leaks of their PII to unauthorized parties and users lose their anonymity due to intense web tracking via cookies [1, 2], device or browser fingerprinting [3–8], cookie synchronization [2, 5, 9–12]. On a direction of more transparent personal data management and user's privacy protection, GDPR from the EU [13] was introduced and enforced a year ago (May 2018) to mitigate or even stop these issues for the EU citizens. GDPR forces websites to stop such activity or receive informed consent from their users and online visitors for any potential tracking and data collection and processing they (1st-parties) may do, and also for any data sharing they may do with other 3rd-parties. Recent studies [14–17] have investigated the aftermath of GDPR, and the effects this privacy legislation had on the online tracking ecosystem, and how websites may have reduced their tracking activity after GDPR was enforced.

However, the web tracking ecosystem is constantly evolving and adapting to blocking methods. Trackers continue their aggressive activity by pushing further, and tracking users across multiple domains, and by employing cookieless, machine learning-based methods to track users across their devices [18–20], all in the name of "more effective retargeting ad-campaigns".

In this present work, we build on and extend previous measurement studies and methods on web tracking, and perform a first of its kind longitudinal study to measure the changes of this ecosystem in the last 2 years, with three different levels of analysis. We do this analysis using 6 crawls of top Alexa websites in time snapshots of a few months apart. The contributions of this work focus on the analysis of the web tracking ecosystem in three levels through time.

First, we look into the network activity that 3rd-parties impose on each publisher (1st-party) hosting them. With this first-level analysis, we confirm existing reports that claim reduction in tracking, by measuring general HTTP network activity from 3rd-parties. However, while less trackers are currently present on the websites (~ 12%), in the last snapshots we measure an increased outgoing network activity to 3rd-parties (e.g., 50% more connections are being established), suggesting potential increase of PII leaks.

Second, we construct bipartite graphs of "publisher-to-tracker" (*PT*), connecting 3rd-parties with their publishers. With this second-level analysis, we employ graph mining tools and metrics such as clustering coefficient, density, degree and betweenness centralities, coreness, etc., to study the graph properties of the 6 constructed bipartite graphs. We find that the structure of the tracking ecosystem with respect to embeddedness in publishers has not changed significantly through time. Popular publishers seem to have both central and non-central trackers in their websites. However, the detected 3rd-parties seem to be embedding themselves in more 1st-parties. Finally, top degree-centrality trackers such as Google's suite (google-analytics, doubleclick, etc.), Facebook, AppNexus, Criteo, etc., dominate the ecosystem in all time snapshots, without losing their market share of publishers. Moreover, we also identify top betweenness trackers such as Twitter and Adobe which are not in the typical top degree tracker list, but have embedded themselves in central positions in the web ecosystem, between different tracker and publisher communities.

Third, we construct a "tracker-to-tracker" graph (*TT*), connecting 3rd-parties who are commonly found in publishers. With this third-level analysis, we construct *TT* pairs from the *PT* graphs, which can reveal potential collaborations between the involved 3rd-parties. We compare these pairs with ground truth data from confirmed data sharing flows of cookie synching (*CS*) pairs, in two different *CS* instances. Proper cookie synching flows between trackers are not easy to get, as they require activity from real users or persona-based automated browsing to trigger the CS mechanism. Interestingly, we measure a high overlap of *CS* and *TT* pairs (~47%-81% when compared to previous ground truth datasets). We propose that such data information flows and sharing can be inferred from the *TT* graphs with reduced cost in deployment and measurements, as they require only web crawling of 1st-parties.

The rest of the paper is organized as follows: Section 2 covers background concepts and related works on the topic of web tracking, methods used, GDPR-related studies, etc. Section 3 presents the different datasets used for the longitudinal analysis. Section 4 presents the first-level analysis on network activity of publishers and trackers across time. Section 5 and Section 6, present the analysis for the second and third level, where we construct and analyze the bipartite graphs (*PT*) and tracker-to-tracker (*TT*) graphs, respectively. Section 7 concludes this study's findings.

2 BACKGROUND AND RELATED WORK

2.1 Web Tracking Studies

Many works have focused on analyzing the web tracking ecosystem, its internal mechanisms and their impact on user's privacy. One of the first studies on web tracking, by Mayer and Mitchell [21], investigated which information is collected by 3rd-parties and how users can be identified. In another work, Roesner et al. [1] studied the different tracking behaviors and measured the prevalence of trackers on the web, and Falahrastegar et al. [22] measured the existence of cookie synching trackers. By collaborating and synchronizing their cookies, 3rd-parties can to have a more completed view of the users' browsing history. On a similar direction Olejnik et al. [9] investigated cookie synching, and found that it is being employed by a large number of trackers, of different organizations. Moreover, Papadopoulos et al. [12] used a heuristic-based mechanism, to detect information exchanged between advertisers through cookie synchronization. They conclude that 97% of the users are exposed to Cookie Synchronization at least once, and that ad-related entities participate in more than 75% of the overall synchronization.

A plethora of studies investigate stateful tracking techniques (i.e., [1, 2, 9, 23]) and stateless techniques, such as browser fingerprinting [3–8]. Acar et al. [5] investigated the prevalence of "evercookies" and the effects of cookie respawning in combination with cookie syncing. Englehardt and Narayanan [2] conducted a large scale measurement study to quantify both stateful and stateless tracking in the web, and Lerner et al. [24] conducted a longitudinal study of 3rd-party tracking behaviors and found that tracking has increased in prevalence and complexity over time.

2.2 Web Tracking as a Graph Model Problem

A number of previous works have analyzed the ecosystem of trackers and 3rd-parties on the web by modeling the activity of trackers into graphs. By studying the graph network properties, one is able to understand the characteristics of the tracking entities, and dissect the ecosystem and its inner mechanisms. In that direction, Kalavri et al. [25] built a 2-mode bipartite graph based on real users traffic logs, and focused their analysis on the communities formed by the graph vertices. Their analysis showed that trackers are well connected to each other, since 94% of them are in the largest connected component. By applying rule based classification and iterative label propagation methods, they were able to classify unknown trackers with high accuracy.

Urban et al. [10] collected various behavioral data from emulated users located in 20 different EU countries and created the "Cookie Synch" graph, which connects 3rd-parties that share information. They reported that the number of trackers and the number of direct syncing connections decreased through time, since fewer 3rd-parties are present in the publisher domains (40% less syncing connections). Also, based on the properties of their graph, they found that the structure of the ecosystem did not change significantly. Similarly, Bashir et al. [11] constructed a cookie-syncing graph, and measured the importance of the different tracking domains (average node degree, connected components, community detection, etc). They also used this graph to emulate different users and quantify their exposure to tracking domains, and the effectiveness of ad-blocking tools.

In the direction of tracking-blocking techniques, Gervais et al. [26], combined different graph metrics to evaluate the effectiveness of various ad-blockers. Their analysis shows that ad-blockers reduce the number of 3rd-parties, that the the "DNT" header does not really have a strong impact on the blocking of HTTP requests, and also that various organizations have a large coverage in most of the popular websites, under different tracking domains.

2.3 GDPR Enforcement and Web Tracking

The General Data Protection Regulation (GDPR) [13], is a regulatory initiative by the European Union (EU) to harmonize data protection laws between its member states. The GDPR specifies under which circumstances personal data may be processed, how the organizations should handle them, and includes several rights of data subjects and obligations for those processing personal data of EU-citizens. For the web, every website (i.e., 1st-party) and content provider (i.e., 3rd-party) must be GDPR compliant in order to serve EU users.

Since this enforcement directly affects the web and the online advertising and tracking ecosystem, a number of recent works have focused on investigating the state of the ecosystem, the evolution of the privacy policies, and their impact on user's privacy. Iordanou et al. [14] collected data from different users across EU, and identified the directions of tracking flows inside EU. They reported that 85% of the tracking flows terminate in servers inside the EU, and that the most sensitive types of user information (based on GDPR) that is being tracked is health, sex orientation and politics.

Degeling et al. [15] quantified the changes of Privacy Policies on the Top-500 sites of the 28 EU countries. They found that in total 85% of the websites have a privacy policy, and that GDPR did not significantly changed the way 3rd-party cookies are used. Dabrowski et al. [16] compared the usage of persistent cookies between EU and USA, and reported that in most of the EU's accessible sites the cookie usage is eliminated (53% of sites did not install a persistent cookie), and 31% of the EU-specific domains dropped cookie usage in the last two years. In the most recent work, Sorenser et.al [17] measured the changes on the presence of 3rd-parties, before and after GDPR enforcement, in 1200 popular sites across EU. According to their study, there were not significant changes in the general state of the web and the tracking ecosystem, and the GDPR had a potential effect only on specific types of websites, (i.e. shopping, travel).

This present study performs a first-of-its kind longitudinal analysis of the web tracking ecosystem over 6 time snapshots and at 3 different levels of analysis. First, it studies network tracking activity of 3rd-parties embedded in publishers from top Alexa websites. It identifies trends through time and also confirms recently reported results on tracking reduction in the ad-ecosystem. On the other hand, it identifies increased outgoing web requests which hint to possible PII leakage behavior. Second, it constructs weighted bipartite graphs connecting publishers to trackers and studies their graph properties with respect to centrality and possible outliers. It identifies that the common suspects of trackers (google, facebook, etc.) are the top in publisher coverage and remain top across time. Finally, it constructs weighted tracker-to-tracker graphs and compares them with established data sharing flows in cookie syncing graphs. It measures and finds high overlap of TT and CS pairs of trackers, pointing to a practical and cheaper alternative to detecting data sharing flows between web trackers than collecting data from real users.

Table 1: Timeline of web crawls using top Alexa lists, and detected 1st and 3rd-party domains.

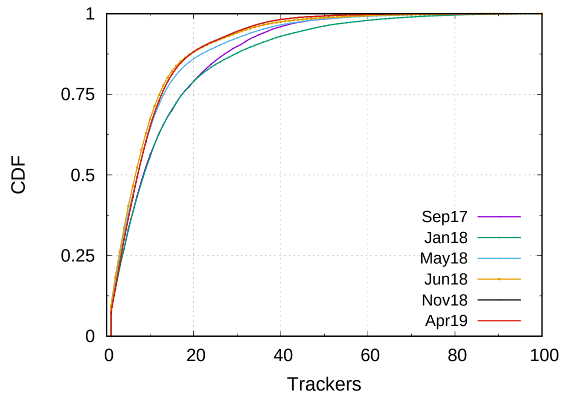
Dataset	Alexa Ranks	1st-parties	3rd-parties
September 2017	10k	8311	848
January 2018	30k	29444	1036
May 2018	80k	73493	1096
June 2018	80k	73813	1068
November 2018	65k	61287	1002
April 2019	65k	59662	819

3 DATA COLLECTION

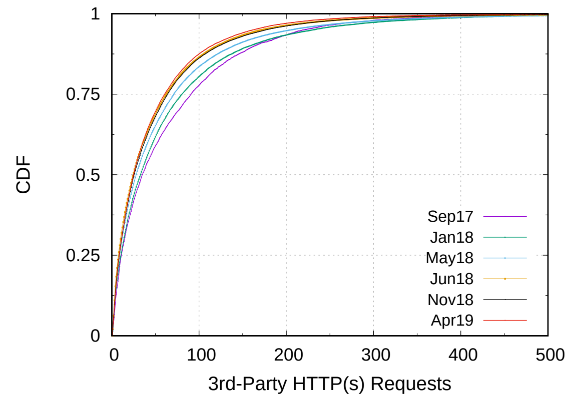
Since our purpose in this work is to conduct a longitudinal study about the evolution of web tracking, and how these practices were affected by the recent EU regulations, we collected and used historical data covering a period of almost 2 years: from September 2017 to April 2019. For collecting an adequate volume of data, we used the OpenWPM framework [2], which enabled us to crawl multiple websites via scripted browsers and store all the HTTP(s) incoming (*GET, PUT, HEAD*) and outgoing (*POST*) requests along with their bodies and headers. We also logged all the cookies that were set by JavaScript, and stored various other crawl-related data (i.e.: time of visit, failed connections, HTML files, etc.).

During the crawling we did not set the "Do Not Track" flag, and we configured our browser to accept all 3rd-party cookies and requests. We used the "bot detection mitigation" technique offered by OpenWPM (i. e., scrolling randomly up and down) and handled the duration of each visit on a website (approximately 10 seconds). Moreover, we empirically set the timeout for a website to respond to 30 seconds.

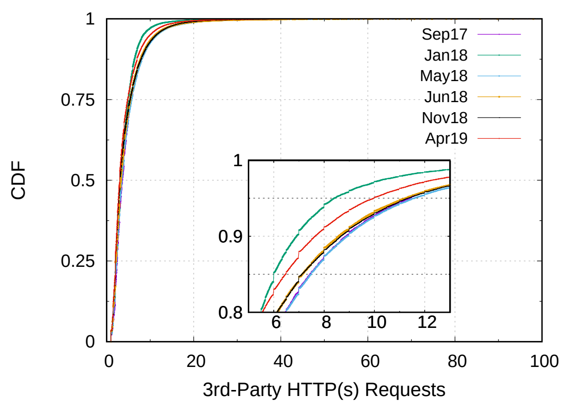
We deployed the framework on a single computer at a European Institution, having a unique IP address, in this



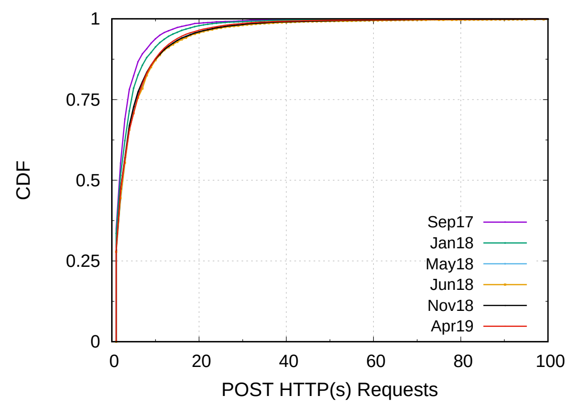
(a) Unique 3rd-parties (domains) embedded by each publisher.



(b) Total number of 3rd-party HTTP(s) Requests per publisher. For this case, we refer to HTTP(s) requests as all the incoming and outgoing connections of the browser during the loading of a publisher.



(c) Average number of 3rd-party HTTP(s) requests per publisher. To compute the average number we sum all the 3rd-party requests (Figure 1b) and divide them with the number of trackers (Figure 1a).



(d) Outgoing (POST) 3rd-party HTTP(s) requests per publisher.

Figure 1: Cumulative Distribution Functions (CDFs) of the measured tracking activity in HTTP(s) requests and tracking domains, through time.

way avoiding any content biases or any type of location based discrimination. As a website input corpus, we took the Alexa Top 1 million list [27], and based on our available resources we crawled each time the top websites covering different subsets of the list. Starting from September 2017, we repeated our crawls approximately every 5 to 6 months. We also performed two consequent crawls just before and after the GDPR enforcement (i.e., mid May 2018 and beginning of June 2018), in order to identify the potential side effects on the ecosystem from the enforcement of this legislation.

In each snapshot of our dataset, we define two different entities: (i) *Publishers*, which are the websites that the users explicitly visit (*i.e., 1st-parties*), and (ii) *Trackers*, which refer to all the *3rd-party* domains that are embedded within those pages (i.e., domains from which resources are fetched, that set cookies, etc.). We use the Disconnect List [28], a popular browser ad-blocking list, in order to identify which requests are directed towards trackers in the HTTP(s) traffic of each website. Since we focus only on 3rd-party tracking, in this study we only consider publishers that embed at least one tracker. For the rest of this paper, and for simplicity, we will interchangeably refer to 1st-parties or publishers, and to 3rd-parties or trackers. A detailed description of our dataset is presented in Table 1.

4 NETWORK ACTIVITY OF TRACKERS

Using the time snapshots collected through the last two years, we perform an analysis of the network activity of trackers that are embedded in the publishers crawled. Figure 1 summarizes different dimensions of the tracking activity of the 3rd-parties present in the publishers of our datasets. The unique number of trackers embedded in each publisher is illustrated in Figure 1(a). In general, there is a clear decrease in the number of trackers through time, since 50% of the publishers until January of 2018 communicated with almost 10 individual trackers in every visit, but this changed by the end of April 2019, when this number was reduced to almost half, with the lowest value measured in June 2018. This might not be a typical trend of the ecosystem, since this "momentary" elimination of trackers could be caused due to the GDPR enforcement in May 2018. However, further investigation is needed towards this direction in the future, to see if publishers go back to the same levels of tracking as before GDPR.

To measure the network activity, we also report the total number of 3rd-party HTTP(s) requests per publisher in Figure 1(b). Clearly, the trend on the distributions is similar: we observe that through time there is a decline in the average number of 3rd-party requests per publisher. Summarizing the previous two measures, in Figure 1(c) we also compute the average number of 3rd-party requests per publisher. In theory, since both number of trackers and total requests are in decline, their ratio (i.e., average) should also be reduced through time. This trend is observed for the majority of websites. However, for a small set of publishers in the last snapshot ($\approx$ 10% of publishers), the average number of HTTP(s) requests is increased.

Moreover, since tracking is connected with PII leakage, we focus on the outgoing (POST) 3rd-party HTTP(s) requests, to capture the cases of potential information sharing. In Figure 1(d), the number of POST requests identified in the network activity of each publisher, is reported. Up to 50% of the publishers trigger around 2 outgoing HTTP(s) requests, which is reasonable considering the type and time of the crawl that we collected the data. We remind the reader that we only visited the main page of each publisher, without having any interaction (e.g., form completion, data typing, etc.) with the inner elements that could potentially trigger any POST request. Interestingly, for the rest 50% of the publishers, there is an increase in the number of POST requests through time, from 2 to 3 requests in the median case, between the first and last snapshots. This behavior of 3rd-parties, initiating more POST requests, may indicate attempts to collect and send more information of the user to external entities.

Previous works on this topic [10, 14, 29], also report similar declines in the number and frequency of trackers. Specifically, in [29] they reported that fewer 3rd-parties are present in specific categories of websites. These facts might be a side effect of the GDPR being enforced, or other effects of the general evolution of the tracking ecosystem.

5 PUBLISHERS & TRACKERS: PT GRAPH

In this section, we use the datasets collected through time to construct bipartite graphs of publishers connected to their trackers (*PT*) (i.e., 3rd-parties that were found in each publisher). We analyze the properties of such graphs when they are unconstrained, i.e., all publishers and trackers available are included (Sec 5.1), or when we fix the set of publishers to the top 10k of Alexa (Sec 5.2).

5.1 Unconstrained PT graphs

Graph Construction. In order to build the graphs from our datasets, we follow a similar approach as introduced by Kalavri et al. [25]. We create a set of *2-mode* graphs of publishers and associated trackers, where the edges of each graph connect vertices of different modes. In this graph, a publisher can connect to multiple trackers, and a tracker can connect to multiple publishers. However, no publishers are connected with each other, and no trackers are connected with each other. We represent all the domains that a browser requests, as a 2-mode graph:

- V_P represents the set of websites (publishers) a user visits.
- V_T represents the set of trackers embedded in publishers.
- E_w is the set of weighted edges connecting vertices of the two different modes.
- $w = (i, j)$ is the weight of the edge connecting tracker i with publisher j.

We move beyond the state-of-art (i.e., [25]) and add weights on the edges, since we want to represent the number of HTTP(s) requests between a publisher and a tracker, with the associated numerical value. The weight w=(i,j), encodes the number of times that a tracker i communicated via HTTP requests with a publisher j.

Data Filtering & Graph Metrics. As we reported in Section 3, in each snapshot we crawled a subset of the Alexa list. Since we want to create a connected representation of the bipartite graphs, we use the Largest Connected Component (LCC) of each graph. The connected component is the subgraph of the total graph in which there exists at least one path between any two of its vertices. In our datasets, there are some isolate groups of nodes which include websites that communicated with one or two different, but not popular trackers; we exclude such isolates. On each of our final graphs the LCCs contain, on average, $\sim$ 95% of publishers and $\sim$ 90% of trackers of the originally crawled lists.

This type of connected graphs allows us to apply various graph metrics, in order to quantify the properties of each one, and also to compare them across time. Building on

Table 2: Full PT graph characteristics. Number of trackers(|T—), Number of vertices (|N—), number of edges (|E|), normalized average weight per edge (*W*), Average vertex clustering coefficient (*CC*), density (*DE*) and diameter (*DD*).

Dataset	—T—	—N—	—E—	W	CC	DE	DD
Sep17	830	8605	105349	0.010	0.019	0.002	6
Jan18	993	29396	408589	0.009	0.013	0.0009	8
May18	1027	72827	813223	0.011	0.0008	0.0003	8
Jun18	923	72805	749034	0.010	0.0009	0.0002	8
Nov18	830	60539	631777	0.010	0.008	0.0003	8
Apr19	781	60101	680378	0.009	0.0009	0.0003	4

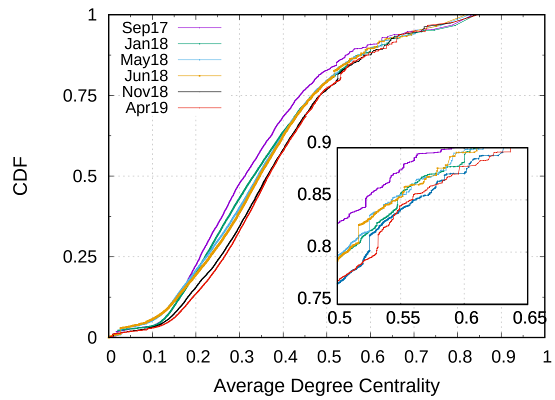


Figure 2: Normalized average degree centrality on the unique set of trackers embedded in each publisher.

the previous works on this topic [10, 11, 25], we use the same metrics such us: Density, Diameter, Average Clustering Coefficient, Degree Centrality, Betweenness Centrality, and Coreness-Periphery (a metric not analyzed before in literature). Detailed definition of these metrics can be found in [30–32]. We study in depth the structure of the tracking ecosystem and its relative changes, on each time snapshot. Finally, the generated graph models and their properties on each of the snapshots are given in Table 2.

5.1.1 How stable are the PT graph properties over time? The differences in the number of nodes and edges between graphs reflect the different number of visited publishers in each snapshot. In general, the average clustering coefficient measures the degree of which the nodes of a graph tend to cluster together (i.e., tend to close triangles between triplets of nodes, or quadruples in bipartite graphs). The low value of this metric on each of the graphs is associated with the bipartite connectivity between the sets of nodes [33]. This metric, in conjunction with the low density, reveals the sparse connections between the different groups of nodes.

Regarding the number of trackers, there is a stable reduction through time, which comes in parallel with the measured elimination in the average number of trackers, as discussed in Section 4. In general, the characteristics and the distance metrics computed on the graphs reveal a consistent structure of the ecosystem during the focal period of our analysis, except for the most recent snapshot where the diameter and average clustering coefficient reach their lower values, while the number of edges increased. This trend captures an increase in the connectivity of the graph nodes, pointing to tracker nodes being closer in the graph, as we also show in the next paragraphs.

Following the previous measures on the structure of the graph, we also compute the *Degree Centrality* for the two sets of vertices in our bipartite graphs. Degree Centrality simply measures the degree of a node, i.e., the number of edges it has attached to it. This metric is often a highly effective measure of the importance of a node, since the nodes with higher degree are clearly more central and potentially can reach more nodes in the graph in 1-hop communications. Figure 2 illustrates the average degree centrality of the trackers, that are embedded in each publisher of our snapshots. This means that for each snapshot, we collect the trackers embedded in each publisher, and compute the average of degree centrality of these trackers. This, in effect, measures how embedded these trackers are, in other publishers.

Across all snapshots, there is a balanced distribution on the degrees of the trackers. In particular, up to the median or 50% of the publishers, they contain trackers with average degree = 0.3-0.35, which means ∼ 2200 other publishers are embedded in the first snapshot, and ∼ 15000 other on the final snapshot. Also, through time there is an increase in the average degree of trackers per publisher. This taken in conjunction with the previous analysis regarding the graph characteristics of the last snapshot (i.e., Apr19, Table 2), points to a transformation in the connectivity of the trackers that are active on the graph, taking on more publishers, and resulting in higher degree centrality.

5.1.2 Do popular publishers prefer popular trackers? Here we quantify if and how the popularity of a website (e.g., via the Alexa Ranking [27]) affects the tracking activity of the embedded 3rd-parties. One could argue that popular trackers will be tempted to collaborate only with popular websites and vice-versa, since popular websites mean more user traffic, and thus more users tracked across the web from a reduced set of top websites. To investigate this argument, we extract the Alexa rankings of the publishers along with the individual degree of each embedded tracker, for each snapshot. Figure 3 illustrates the distribution of tracker degrees of each subset of publishers, according to Alexa Rank for

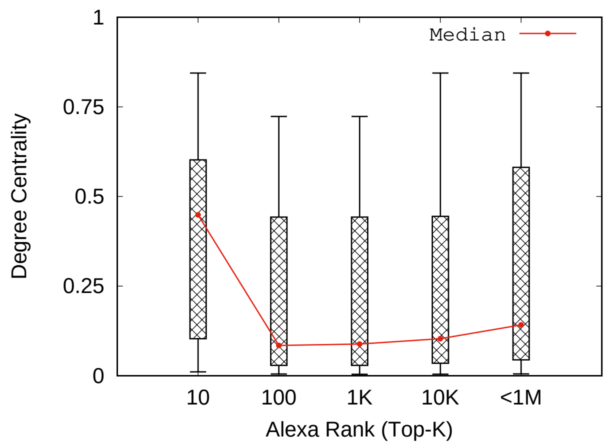


Figure 3: Distribution of the degree centrality of 3rd-parties, as measured on each subset of publishers sorted by Alexa Rank, on the November 2018 snapshot.

the period of November 2018 (similar distributions were extracted for all snapshots and are excluded for brevity). From these results, the previous argument is not supported, since the distribution of the median and all values is similar across the Alexa rankings. This outcome reveals that well-known publishers collaborate with both popular trackers as well as with smaller ones, allowing user data to reach diverse parts of the tracking ecosystem. It also validates the fact that top trackers such us Google and Facebook have a coverage of more that 70% of the popular websites [34].

5.2 Common publishers PT graphs

Going one step further, we want to extract insights about the structure of the tracking ecosystem and its inner properties. As we detailed in Table 1, our datasets are formed by different number of websites belonging to different popularity rankings. Considering the fact that Alexa list is volatile and is constantly changing [35], we focus on a common subset of publishers. This subset contains only those publishers that are found across all time snapshots, i.e., 5100 publishers. Focusing on this immutable set of publishers, we create new PT' graphs and measure the evolution of the state of the tracking ecosystem.

We construct similar representations of Bipartite Graph models that we introduced in Section 5.1, and apply the same filtering on the LCC to construct connected graphs. On these graphs, the LCCs contain approximately the 99.8% of the original set of publishers. The overview of the generated graphs can be found in Table 3.

Similarly to § 5.1, there is a diverse set of nodes and edges due to the different number of trackers that each publisher communicates with. Compared to the previous graphs, these subsets are more dense, a fact that is also reflected in the average clustering coefficient.

Table 3: Characteristics of bipartite PT graphs for common publishers.

Dataset	—N—	—E—	W	CC	DE	DD
Sep17	5710	74037	0.0013	0.024	0.022	7
Jan18	5688	86875	0.0012	0.024	0.027	7
May18	5678	81717	0.0013	0.021	0.026	8
Jun18	5654	76077	0.0013	0.023	0.025	7
Nov18	5636	71481	0.0013	0.026	0.024	8
Apr19	5602	72722	0.0011	0.022	0.025	4

5.2.1 How do publishers & trackers compare in centrality? We measure the degree centrality, betweenness centrality and coreness periphery of each node in our bipartite graphs, in order to quantify the backbone structure of each graph in terms of connectivity and centrality of nodes. Figures 4, 5, 6, 7, 8, 9 illustrate the measured values of each metric as computed on each snapshot for publishers and trackers.

Focusing on Figures 4 and 7, in total the degree centrality for publishers is an order of magnitude smaller compared to the one for trackers. Also, 50% of the publishers' degree centrality is ≤ 0.025, with only $\sim 5\%$ of the publishers having more than 0.01. Conversely, tracker nodes have higher degree centrality scores, with $\sim 5\%$ measured at ≥ 0.15, hinting to the fact of the well-known trackers that cover approximately all the publishers (e.g.: google, facebook, etc.). Some example of publishers with highest degree centrality are telegraph.co.uk , newyorker.com and rollingstone.com, and examples of trackers with high degree centrality include google-analytics.com, criteo.com, and facebook.com. By definition, since we are analyzing a bipartite network constructed by two set of nodes of different sizes and edge weights (the same trackers are communicating with various publishers at a different rate), it is reasonable for the tracking nodes to be more central in the network structure.

Regarding the betweenness centrality, the measured scores are given in Figures 5 and 8 for publishers and trackers, respectively. Publishers' low scores are expected due to the fact that betweenness centrality measures the extent to which a node lies on paths between other nodes, and publishers are not connected to each other but only with other trackers. On the other hand, trackers that are highly active, central and well known, are found on the tail of the distribution, with scores ≥ 0.02. Some example of publishers with highest betweenness are starbucks.com , livescore.com and

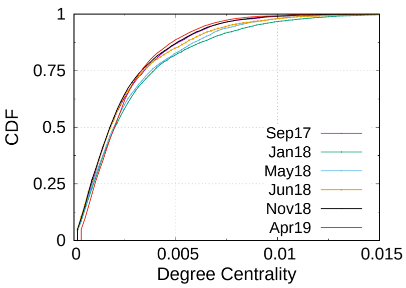


Figure 4: Degree centrality of publishers for the common set of publishers PT graphs.

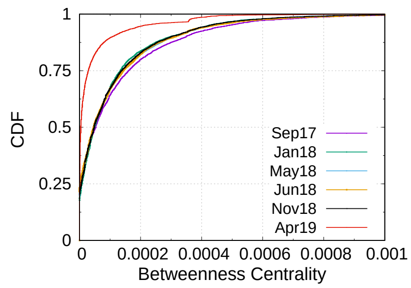


Figure 5: Betweenness centrality of publishers for the common set of publishers PT graphs.

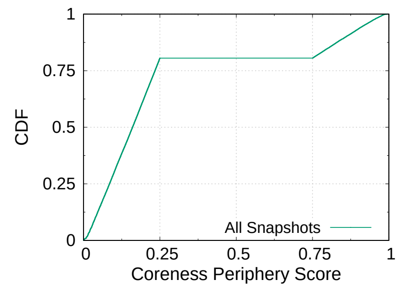


Figure 6: Coreness Periphery of publishers for the common set of publishers PT graphs.

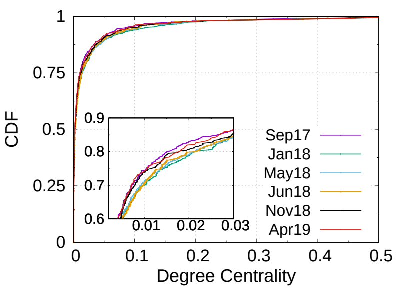


Figure 7: Degree centrality of trackers for the common set of publishers PT graphs.

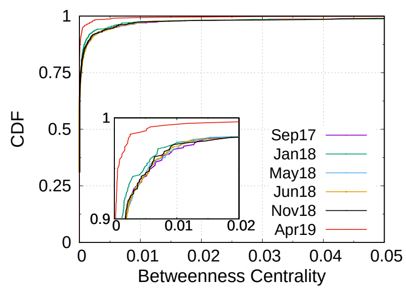


Figure 8: Betweenness centrality of trackers for the common set of publishers PT graphs

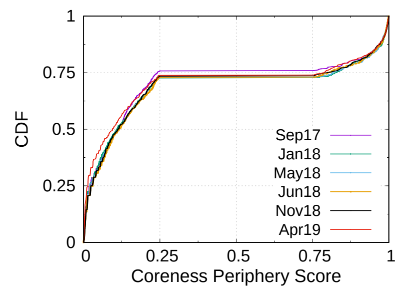


Figure 9: Coreness Periphery of trackers for the common set of publishers PT graphs.

cnn.com, and accordingly the set of trackers with high betweenness includes moatads.com, instagram.com and scorecardresearch.com.

Finally, when comparing coreness scores, in Figures 6 and 9 for publishers and trackers, respectively, we note that trackers tend to occupy positions in the graph with higher coreness than publishers. Interestingly, we observe the two classes of nodes in each graph, with periphery nodes being in the beginning of each CDF (up to ~0.25 coreness score), and core nodes being in the end of each CDF, with coreness score > 0.75). Well known publishers such us: sfgate.com and sport.es and indianexpress.com have high coreness periphery scores. Finally some of the highest coreness score trackers are yandex.ru, adroll.com and amazon-adsystem.com.

5.2.2 How do these centrality metrics correlate? After studying the inner structure of the bipartite networks, we evaluate the relationship between the centrality of nodes for these three metrics. This study will help us understand if the nodes (trackers or publishers) tend to be top (or bottom) in all metrics at the same time, or if there is some disassociation between these metrics. Such disassociation can reveal outliers of nodes who are high in one metric but low in another.

Table 4: Pearson Correlation Coefficient between the distribution of: Degree Centrality & Betweenness Centrality (*DC-BC*), Degree Centrality & Coreness Periphery (*DC-CP*), and Betweenness Centrality & Coreness Periphery (*BC-CP*), for each set of the PT graph for common publishers across snapshots. The maximum confidence level (*p-value*) across all measures was 0.009. The publishers are ranked according to the Alexa list, while trackers according to their Degree Centrality.

Dataset	Publishers			Trackers		
	~DC-BC	*~DC-CP*	*~BC-BC*	*~DC-BC*	*~DC-CP*	*~BC-CP*
Sep17	0.50	0.69	0.26	0.75	0.47	0.26
Jan18	0.43	0.83	0.3	0.73	0.50	0.248
May18	0.47	0.75	0.20	0.73	0.50	0.248
Jun18	0.46	0.74	0.24	0.75	0.48	0.25
Nov18	0.42	0.72	0.19	0.77	0.49	0.26
Apr19	0.43	0.71	0.23	0.54	0.47	0.18

For this reason, we compute the Pearson Correlation score between the distributions of Degree Centrality, Betweenness Centrality and Coreness Periphery for the trackers and publishers, independently. A detailed report on the correlation scores is given in Table 4. In general, there is a strong association between all the distributions with high confidence level. Trackers have higher correlation score in the Degree-Betweenness comparison, hinting their importance and central role on the network, regardless of metric used. On the contrary, the correlation scores for Degree-Coreness for the publishers are measured higher than trackers, again validating the importance of the publishers of this union set who tend to have high and important position in the network structure. Finally, across both types of nodes, Betweenness-Coreness correlation scores are lower, pointing to a disassociation between the two measures.

5.3 What is the ecosystem's current state?

After analyzing the inner properties of the graph models, we want to qualify the importance of the top tracking nodes, for the ecosystem and it's actual extent to user's privacy. To have a clear view of the most important trackers at each time of our crawl, we measured the degree centrality of the trackers, and ranked them accordingly. A complete report about the Top-25 trackers across time is given in Table 5.

Interestingly, the Top-22 trackers remained the same across all snapshots with minor fluctuations on the internal ranking. The most important trackers contain the "big", well-known entities of the ad-industry, such us Google, Facebook, Twitter and Criteo, as well as some smaller but still established companies such as Bluekai and Taboola. The table also reports those trackers that gradually climbed into the Top list, illustrating in this way the plurality of the web tracking ecosystem.

Furthermore, in Table 6 we make a similar investigation but for betweenness centrality. We note that the top of the list is populated by similar trackers. However, new entities such has linkedin.com, ads-twitter.com and veresttech.net emerge, which demonstrate a central position in the ecosystem with respect to mediating flows between distant parts in the ecosystem.

Overall, the almost immutable list of top trackers in either of the two metrics points to the fact that the GDPR enforcement had no effect on them either in their importance in the web tracking ecosystem, or their coverage across websites. Also, in the previous sections we found that trackers are present in more websites (as time passed by), but at the same time web requests have been reduced. We can conclude that there may be a "shift" of publishers on the type of business relationships they make with the well-known, less privacy intrusive, and GDPR-compliant trackers.

Table 5: Top-25 Trackers ranked by Degree Centrality and labeled under the umbrella of company/organization and the average percentage of coverage in publishers through time.
(*): Set of trackers that form the Top-25 list across all snapshots.
(+/-): Set of trackers that were part of the Top-25 in one or more snapshots, but their rank decreased through time below Top-25.
(+): Set of trackers that were not part of the Top-25 in the first snapshot, and they climbed in the Top-25 through time.

Tracker	Organization	Publishers(%)
(*) google-analytics.com	Google	81.0
(*) doubleclick.net		70.0
(*) google.com		51.0
(*) googleapis.com		57.5
(*) googletagmanager.com		36.5
(*) facebook.com	Facebook	44.5
(*) facebook.net		41.5
(*) googletagservices.com	Google	28.0
(*) gstatic.com		44
(*) googlesyndication.com		28.3
(*) googleadservices.com		19.0
(*) cloudfront.net	Amazon	18.0
(*) adnxs.com	App Nexus	18.0
(*) criteo.com		13.0
(*) criteo.net	Criteo	13
(*) scorecardresearch.com	comScore	12.5
(*) twitter.com	Twitter	20.0
(*) rubiconproject.com	Google	12.5
(*) pubmatic.com	Pubmatic	11
(*) openx.net	OpenX	8.5
(*) casalemedia.com	Casale Media	9.0
(*) advertising.com	Verizon Media	7.0
(+/-) quantserve.com	Quantcast	9.0
(+/-) adsrvr.org	The Trade Desk	9.0
(+/-) taboola.com	Taboola, Inc	7.0
(+/-) nr-data.net	New Relic	8.0
(+/-) 2mdn.net	Google	6.0
(+/-) bluekai.com	BlueKai	8.0
(+) alexametrics.com	Amazon	4.0
(+) demdex.net	Adobe	7.0
(+) newrelic.com	New Relic	8.0

6 TRACKER TO TRACKER: TT GRAPH

We continue our exploration in studying the web tracking ecosystem using another type of graph, in which we connect trackers to other trackers, based on the common publishers they were found on. The intuition for the construction of this graph is that such tracker pairs may indicate frequent collaborations between web trackers, and may point

Table 6: Top-25 Trackers ranked by Betweenness Centrality (BC) and labeled under the umbrella of company/organization, along with their BC score and percentage of coverage in publishers by April 2019. We highlight the trackers not present in Table 5.

Tracker	Organization	BC	% Publishers
googletagmanager.com	Google	0.077	46.55
doubleclick.net		0.012	72.15
googleadservices.com		0.008	19.59
googletagservices.com		0.006	28.82
gstatic.com		0.005	44.75
cloudfront.net	Amazon	0.005	14.52
newrelic.com	New Relic	0.004	7.63
rlcdn.com	Live Ramp	0.003	8.66
pubmatic.com	Pubmatic	0.003	13.1
google.com	Google	0.030	59.09
nr-data.net	New Relic	0.002	7.61
facebook.com	Facebook	0.002	44.01
everesttech.net	Adobe	0.002	3.52
casalemedia.com	Casale Media	0.002	7.48
alexametrics.com	Amazon	0.002	4.46
ads-twitter.com	Twitter	0.002	4.72
adsrvr.org	Trade Desk	0.002	5.67
adnxs.com	App Nexus	0.002	13.83
twitter.com	Twitter	0.001	13.32
rubiconproject.com	Google	0.001	10.15
quantcount.com	Quantcast	0.001	4.11
openx.net	OpenX	0.001	8.59
linkedin.com	Microsoft	0.001	4.71
advertising.com	Verizon Media	0.001	7.23

to potential data sharing among them. We focus again on the data used in Section 5.2, with the subset of publishers who are common across all snapshots (5100 publishers). In this way, we control the publishers used, and allow the trackers who are linked to these publishers to change.

6.1 TT graph construction

In this section, we build tracker-to-tracker graphs, who are undirected but weighted, $TT = (V_{TT}, E_{TT})$, and originate from their corresponding PT graphs:

- V_{TT} represents the set of trackers embedded in publishers.
- E_{TT} is the set of weighted edges connecting two trackers, if and only if both trackers coexist in at least two different publishers.
- $w = (i, j)$ is the weight of the edge connecting tracker i with tracker j.

The weight *w=(i,j)*, encodes the number of publishers that tracker i and tracker j coexisted. A detailed description of our *TT* graph characteristics and their distance metrics is presented in Table 7.

Table 7: Characteristics of TT graphs produced from the PT graphs with common publishers across all snapshots.

Dataset	—N—	—E—	W	CC	DE	DD
Sep17	815	63177	0.004	0.69	0.19	4
Jan18	774	53325	0.006	0.67	0.17	4
May18	846	76686	0.005	0.72	0.21	4
Jun18	824	69145	0.005	0.70	0.20	4
Nov18	834	72265	0.005	0.71	0.20	4
Apr19	841	74012	0.005	0.71	0.20	4

The *TT* graph of each snapshot has a fairly dense structure (average Density $0.19 - 021$). The number of nodes as well as the number of edges, are comparable through snapshots which is reasonable since we focused only on the common publishers in the dataset, and extracted the trackers that were present in each snapshot. Interestingly, in May 2018 (before GDPR) the number of edges reached a maximum which was not surpassed in subsequent snapshots. In general, in all TT graphs trackers are well connected and clustered with each other (average clustering coefficient $0.67-0.72$). These properties of the TT graphs highlight the dense structure of the tracking ecosystem, and how the 3rd-parties potentially share user's information.

6.2 Is cookie synchronization pairs present in TT graphs?

Since the purpose of the TT graph construction is to study potential data sharing among web trackers, we need to compare the constructed pairs with existing data that already measure such data sharing flows among web trackers. Such ground truth data are called cookie synchronizations and we received access to two such datasets provided by Papadopoulos et. al. [12] and Bashir et al. [11]. These datasets contain pairs of 3rd-parties that performed Cookie Synchronizations while real users [12], or crawlers [11] were browsing the web. Interestingly, the dataset from [12] also includes a normalized frequency on each pair, encoding the number of times the two entities of the pair shared information. Following a similar representation as with the TT graphs, we create 2 undirected CS graphs $CS = (V_{CS}, E_{CS})$, with weighted edges for the data from [12], i.e.: *w(i,j)*= number of times that pair of trackers (i,j) performed information exchange (cookie-syncing). The first CS graph from [12] has 4656 trackers and 8582 edges connecting them, whereas the second CS graph from [11] has 59 trackers and 200 edges connecting them.

Table 8: **Percentage of overlap between the different sets of trackers for the CS edges extracted from [12]. We refer to the common number of trackers as $|N|$.**

Dataset	$\|N\|$	$\|E_{CS}\|$	$\|E_{TT}\|$	O_{common}	$O_{\neg CS}$	$O_{\neg TT}$
Sep17	226	3015	28631	59.70	49.10	1.70
Jan18	226	3015	28631	59.70	49.10	1.70
May18	226	3024	27353	58.10	46.80	1.90
Jun18	222	3003	24943	55.30	44.70	2.30
Nov18	214	2820	20976	52.20	39.90	2.20
Apr19	210	2929	18669	47.30	37.90	3.30

For the investigation of existence of CS pairs into TT pairs, we define the following sets of pairs of trackers:

- E_{CS}: set of edges in a CS graph (i.e., pairs of trackers who have performed CS at some point in time).
- E_{TT}: set of edges in a TT graph (i.e., pairs of trackers who have co-existed in at least 2 publishers).
- $\neg CS$: set of non-edges in a CS graph (i.e., pairs of trackers who have not performed CS at any point in time).
- $\neg TT$: set of non-edges in a TT graph (i.e., pairs of trackers who have not co-existed in any publishers)

We also define the following overlaps of the above sets:

- $O_{common} = E_{CS} \cap E_{TT}$
- $O_{\neg CS} = \neg CS \cap E_{TT}$
- $O_{\neg TT} = E_{CS} \cap \neg TT$

To have an accurate measurement between the different overlaps of the previously defined sets, on each TT snapshot we filter the edges, and store only those that are parts of the common trackers between each CS and each TT graph. A detailed report on the percentages of overlap between the different sets for each of the two CS graphs, is given in Tables 8 and 9 for the two CS datasets.

According to Table 8, the overlap between CS and TT edges across snapshots is $47-60\%$. In the smaller CS dataset, as shown in Table 9, this overlap is even higher, ranging to $64-81\%$. Considering that the TT graphs were built artificially using the combination of trackers of each dataset as they appeared on publishers, this high overlap gives us an indication about the "nature" of CS pairs, and how such data sharing flows can be found in a TT graph. This a crucial finding: we can detect potentially collaborating pairs of trackers who may be sharing data of users, without the need to deploy infrastructure to collect real users' data, or train artificial personas to collect CS activity.

Moreover, since the first CS graph was weighted, we investigated how well the TT edges that overlap with the CS edges cover the distribution of weights. That is, how representative are the TT edges of the CS edges, with respect to weights (i.e., intensity of communication between trackers). We found that the common TT edges cover well the distribution of the CS weights as well. Finally $\sim 2\%$ of the overlapping TT edges are edges with high frequency on the CS graph.

Table 9: **Percentage of overlap between the different sets of trackers for the CS edges extracted from [11]. We refer to the common number of trackers as $|N|$. The $O_{\neg TT}$ value was measured 0% across all TT graphs.**

Dataset	$\|N\|$	$\|E_{CS}\|$	$\|E_{TT}\|$	O_{common}	$O_{\neg CS}$
Sep17	42	104	7321	80.80	76.40
Jan18	41	102	7169	80.40	75.90
May18	41	104	6860	78.90	71.80
Jun18	41	104	6389	73.10	69.20
Nov18	40	104	5896	69.30	68.50
Apr19	39	104	5309	64.40	63.20

7 DISCUSSION AND CONCLUSION

The present work performed a first of its kind longitudinal study and measured the changes of the web tracking ecosystem in the last 2 years, and by employing three different levels of analysis. The analysis through time was performed using 6 crawls of top Alexa websites in time snapshots of a few months apart.

In the first level of analysis, we focused on network-level traffic of trackers to publishers with the following findings:

- There are fewer trackers embedded in the websites through time. We measured a reduction of 12% for the median case and 25% for the 90% percentile case.
- There are fewer HTTP(s) requests directed to 3rd-parties through time. We measured a reduction of 17% for the median case and 13% for the 90% percentile case.
- There is a higher potential for PII leakage, since the number of POST requests significantly increased through time. We measured an increase of 50% for the median and the 90% percentile case.

In the second level of analysis, we constructed bipartite graphs of publishers connected with their trackers (PT graphs) and studied their graph properties through time. We identified top trackers and how they are employed from various types of websites, regardless of Alexa ranking. In summary, we made the following findings:

- The network structure of the tracking ecosystem and how trackers are embedded in publishers remained the same through time.

- The same 3rd-parties that existed through time have been forced to cover more websites, and especially the top central trackers.
- Popular websites collaborate and communicate with both well connected or "popular", as well as "uncommon" trackers of the ecosystem.
- Top trackers in terms of publisher coverage and centrality in the PT graphs (e.g., google-analytics, doubleclick, facebook, criteo, appnexus, etc.), remained top across time.
- In terms of node importance on the graph representation of the web tracking ecosystem, trackers appeared to be central in many of the examined centrality metrics (degree, betweenness, coreness in the graph).
- GDPR enforcement had no effect on the "big" tracking entities of the web ecosystem.

Finally, in the third and deepest level of analysis, we constructed tracker-to-tracker graphs (TT) for trackers who coexisted on the same publishers. We compared these TT graphs with confirmed cookie synchronization (CS) pairs of trackers. These CS pairs are established ground truth information flows between tracking entities. We found high overlap between the TT edges and the CS pairs. This means we can detect potential cookie syncing activity and data sharing flows between trackers with a practical and cheaper alternative than collecting data from real users. Furthermore, the investigation of the TT graph properties uncovered "hidden", and not studied so far, relationships between the CS and TT pairs. Therefore, our study points to the future opportunity for building an accurate and representative graph model of the web tracking ecosystem, without the overhead of running complex data collection techniques or requiring user consent.

REFERENCES

[1] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, NSDI'12, pages 12–12, Berkeley, CA, USA, 2012. USENIX Association.

[2] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 1388–1401, New York, NY, USA, 2016. ACM.

[3] Peter Eckersley. How unique is your web browser? In *Proceedings of the 10th International Conference on Privacy Enhancing Technologies*, PETS'10, pages 1–18, Berlin, Heidelberg, 2010. Springer-Verlag.

[4] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens, and Bart Preneel. Fpdetective: Dusting the web for fingerprinters. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, pages 1129–1140, New York, NY, USA, 2013. ACM.

[5] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 674–689, New York, NY, USA, 2014. ACM.

[6] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proceedings of the 2013 IEEE Symposium on Security and Privacy*, SP '13, pages 541–555, Washington, DC, USA, 2013. IEEE Computer Society.

[7] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. Privaricator: Deceiving fingerprinters with little white lies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 820–830, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[8] Andriy Panchenko, Fabian Lanze, Jan Pennekamp, Thomas Engel, Andreas Zinnen, Martin Henze, and Klaus Wehrle. Website fingerprinting at internet scale. In *NDSS*, 2016.

[9] Lukasz Olejnik, Tran Minh-Dung, and Claude Castelluccia. Selling off privacy at auction. In *Network and Distributed System Security Symposium (NDSS)*, 2014.

[10] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. The unwanted sharing economy: An analysis of cookie syncing and user transparency under gdpr. *arXiv preprint arXiv:1811.08660*, 2018.

[11] Muhammad Ahmad Bashir and Christo Wilson. Diffusion of user tracking data in the online advertising ecosystem. *Proceedings on Privacy Enhancing Technologies*, 2018(4):85–103, 2018.

[12] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference*, pages 1432–1442. ACM, 2019.

[13] Council of the European Union , European Parliament. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.

[14] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. Tracing cross border web tracking. In *Proceedings of the Internet Measurement Conference 2018*, pages 329–342. ACM, 2018.

[15] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. *arXiv preprint arXiv:1808.05096*, 2018.

[16] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. Measuring cookies and web privacy in a post-gdpr world. In *International Conference on Passive and Active Network Measurement*, pages 258–270. Springer, 2019.

[17] Jannick Kirk Sørensen and Sokol Kosta. Before and after gdpr: The changes in third party presence at public and private european websites. In *WWW '19 Companion Proceedings of the The Web Conference 2019*, United States, 1 2019. Association for Computing Machinery.

[18] Sebastian Zimmeck, Jie S. Li, Hyungtae Kim, Steven M. Bellovin, and Tony Jebara. A privacy analysis of cross-device tracking. In *26th USENIX Security Symposium*, USENIX Security 17, pages 1391–1408. USENIX Association, 2017.

[19] Justin Brookman, Phoebe Rouge, Aaron Alva, and Christina Yeung. Cross-device tracking: Measurement and disclosures. *Proceedings on Privacy Enhancing Technologies*, 2017.

[20] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. Talos: An automated framework for cross-device tracking detection. In *22nd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2019)*, 2019.

[21] Jonathan R. Mayer and John C. Mitchell. Third-party web tracking: Policy and technology. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP '12, pages 413–427, Washington, DC, USA, 2012. IEEE Computer Society.

[22] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. Tracking personal identifiers across the web. In *International Conference on Passive and Active Network Measurement*, pages 30–41. Springer, 2016.

[23] Zhonghao Yu, Sam Macbeth, Konark Modi, and Josep M. Pujol. Tracking the trackers. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 121–132, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

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

[25] Vasiliki Kalavri, Jeremy Blackburn, Matteo Varvello, and Konstantina Papagiannaki. Like a pack of wolves: Community structure of web trackers. In *International Conference on Passive and Active Network Measurement*, pages 42–54. Springer, 2016.

[26] Arthur Gervais, Alexandros Filios, Vincent Lenders, and Srdjan Capkun. Quantifying web adblocker privacy. In *European Symposium on Research in Computer Security*, pages 21–42. Springer, 2017.

[27] Alexa. The top 500 sites on the web. https://www.alexa.com/topsites/category/Top/, 2018.

[28] Disconnect. Disconnect lets you visualize and block the invisible websites that track your browsing history. https://disconnect.me/, 2018.

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

[30] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[31] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[32] Puck Rombach, Mason A Porter, James H Fowler, and Peter J Mucha. Core-periphery structure in networks (revisited). *SIAM Review*, 59(3):619–646, 2017.

[33] Matthieu Latapy, Clemence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large affiliation networks/bipartite graphs. *arXiv preprint cond-mat/0611631*, 2006.

[34] Iskander Sanchez-Rola and Igor Santos. Knockin' on trackers' door: Large-scale automatic analysis of web tracking. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 281–302. Springer, 2018.

[35] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D Strowes, and Narseo Vallina-Rodriguez. A long way to the top: significance, structure, and stability of internet top lists. In *Proceedings of the Internet Measurement Conference 2018*, pages 478–493. ACM, 2018.