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

Konstantinos Solomos*, Panagiotis Ilia*, Sotiris Ioannidis†, Nicolas Kourtellis‡

*University of Illinois at Chicago, USA; †FORTH,Greece; ‡Telefonica Research, Spain

*{ksolom6,pilia}@uic.edu; †sotiris@ics.forth.gr; ‡nicolas.kourtellis@telefonica.com

Abstract—**Websites are constantly adapting the methods used, and intensity with which they track online visitors. However, the wide-range enforcements of regulations such us GDPR and e-Privacy has forced websites serving EU-based online visitors to eliminate, or at least reduce, such tracking activity, given they receive proper user consent. Thus, it is important to analyze the aftermath of such policies, record the evolution of this tracking activity, and assess the overall "privacy health" of the Web ecosystem. This work makes a significant step towards this direction. In this paper, we analyze the ecosystem of 3rd-parties embedded in top websites, which amass the majority of online tracking, through six time snapshots taken every few months apart, in the duration of the last few 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, embeddedness of trackers in publishers, popularity and mixture of trackers across publishers. In the last level of analysis, we dig deeper and look into the interconnectivity of trackers, and how this relates to potential cookie synchronization activity.**

I. INTRODUCTION

Online users' privacy is constantly violated by leaks of their PII to unauthorized parties. Users lose their anonymity due to web tracking via cookies [1], [2], device or browser fingerprinting [3]–[5] and cookie synchronization [6]–[10]. In the last few years, different legislations and directives such as the e-Privacy [11], the General Data Protection Regulation (GDPR) [12] in EU, and the California Consumer Privacy Act (CCPA) [13] in USA, were introduced in an effort to increase transparency in user tracking and help users with personal data management and privacy protection. In particular, GDPR forces websites to stop such activity or receive informed consent from their online visitors for any potential tracking, data collection and processing they may do, and also for any data sharing they may do with 3rd-parties.

Recent studies [14]–[18] have investigated the aftermath of GDPR and its effects on the online tracking ecosystem, and how websites may have reduced their tracking activity. However, the web tracking ecosystem is constantly evolving and adapting to new blocking methods. Trackers continue their aggressive activity across multiple domains, at times unbalanced depending on the type of website [19], and even

employ cookie-less, machine learning-based methods to track users across their different devices [20]–[22], all in the name of "more effective ad-campaigns".

In this work, we build on previous studies and methods on web tracking, and perform a first of its kind longitudinal study to measure this ecosystem's changes over the last few years. We perform this analysis in three levels through time, using six crawls of top Alexa websites in time snapshots of a few months apart. 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 generic HTTP network activity from 3rd-parties. Second, we construct "publisher-to-tracker" (PT) bipartite graphs, connecting 3rd-parties with their publishers. With this second level of analysis, we employ graph mining tools and metrics such as clustering coefficient, density, degree centrality, coreness, etc., to study the graph properties of the six bipartite graphs. We find that the structure of the tracking ecosystem with respect to embeddedness in publishers has not changed significantly through time, and that 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 list but have embedded themselves in central positions in the web ecosystem.

Third, we construct a "tracker-to-tracker" graph (TT), connecting 3rd-parties that are commonly found in publishers. With this analysis, we construct TT pairs from the PT graphs that can reveal potential collaborations between 3rd-parties. We compare these pairs with ground truth data from confirmed data sharing flows of cookie syncing (CS) pairs, in two different CS datasets from past studies. Proper cookie syncing 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 identify a high overlap between CS and TT pairs (∼47%-81% when compared to previous ground truth CS datasets). We propose that such information flows can be inferred from the TT graphs with reduced cost in deployment and measurements, as they require only web crawling of 1st-parties.

II. Background and Related Work

A. Web Tracking & Graph Modeling

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 [23], investigated which information is collected by 3rd-parties and how users can be identified. Roesner et al. [1] studied the various tracking behaviors and measured the prevalence of trackers while Falahrastegar et al. [24] measured the existence of cookie synchronization trackers. Papadopoulos et al. [10] used a heuristic-based mechanism to detect information exchanged between advertisers and found that 97% of the users are exposed to cookie syncing at least once, and that ad-related entities participate in more than 75% of the overall cookie synchronization.

A plethora of studies investigate stateful tracking techniques (e.g., [1], [2], [6], [25]), and stateless techniques such as browser fingerprinting [3]–[5], [8], [26], [27]. Acar et al. [8] 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. [28] conducted a longitudinal study of 3rd-party behaviors and found that tracking has increased in prevalence and complexity over time.

In general, 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. [29] built a 2-mode bipartite graph based on real user 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. Urban et al. [30] collected behavioral data from emulated users located in 20 EU countries and created a cookie synchronization graph that 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. [7] constructed a cookie syncing graph and associated the graph metrics with the existence of different tracking domains.

B. GDPR Enforcement and Web Privacy

The GDPR [12] is a regulatory initiative by the EU to harmonize data protection laws between its member states, and specifies under which circumstances personal data may be processed. Since the regulation directly affects the web and the online advertising and tracking ecosystem, recent works have focused on investigating the state of the ecosystem, the evolution of the privacy policies and their impact on user's privacy from different perspectives [16], [18], [31]. Iordanou et al. [14] collected data from users across EU and identified

TABLE I: Timeline of web crawls using top Alexa lists, and number of unique 1st- and 3rd-party domains we detected in each crawl. The (*) shows which snapshots were collected after the enforcement of GDPR. The overlap row indicates the intersection of common domains across all snapshots.

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
Overlap	**10K**	**5100**	**527**

the directions of tracking flows inside EU. They reported that 85% of the tracking flows terminate in servers located inside the EU, and identified 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 change the way 3rd-party cookies are used. In their most recent work, Sorenser et. al. [17] measured the changes on the presence of 3rd-parties, before and after GDPR enforcement, in 1200 popular websites across EU. Their study shows that there were not significant changes in the general state of the web, and that GDPR had a potential effect only on specific types of websites.

III. Data Collection

Since our purpose is to conduct a longitudinal study on the evolution of tracking, and how it was affected by the recent EU regulations, we collected historical data covering a period of almost 2 years (September 2017 to April 2019). For collecting these data we used the OpenWPM framework [2] to crawl multiple websites via scripted browsers, and we stored all the HTTP(s) incoming and outgoing requests. We also logged the cookies set by JavaScript, and stored various other crawl-related data (i.e., time of visit, HTML files, etc.). During the crawling, we did not set the "Do Not Track" flag, we configured our browser to accept all connections, and we empirically set the timeout for a website to respond to 30 seconds. For not introducing any unnecessary complexity and overhead during the crawls, we do not perform any browsing to visit deeper than the main page of each website. We deployed the framework on a single computer at an EU academic institution, having a unique IP address, thus, avoiding any content biases or any type of location based discrimination. As a website input corpus, we used the Alexa Top 100K list [32], 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 consecutive crawls just before and after the GDPR enforcement, i.e., mid May 2018 and beginning of June 2018.
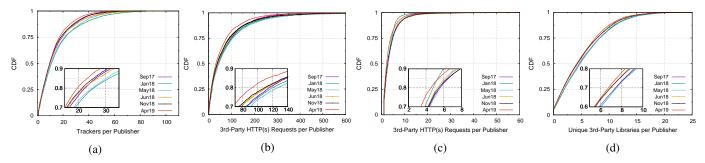
Fig. 1: These figures, from left to right, show (a) the number of unique 3rd-party trackers, (b) the total number of 3rd-party HTTP(s) requests, (c) the average number of 3rd-party HTTP(s) requests, and (d) the number of embedded 3rd-party libraries.

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 the *3rd-party* domains that are embedded within the visited pages (i.e., domains from which resources are fetched, that set cookies, serve content, etc.). We use the most recent, by the time of each crawling, Disconnect List [33] (a popular browser ad-blocking list), in order to identify which requests are directed towards trackers. In this way, we were able to accurately detect the presence of trackers in each page and to identify their behaviors. Since tracking domains change frequently, and only a subset of them might remain constant through time, the list is constantly updated to capture such changes. Furthermore, we only consider publishers that embed at least one tracker in their page. For simplicity, we interchangeably refer to 1st-parties or publishers, and to 3rd-parties or trackers in the rest of the paper. A detailed description of our datasets is presented in Table I.

IV. NETWORK ACTIVITY OF TRACKERS

Using the aforementioned six time snapshots, we perform an analysis on the network activity of trackers that are embedded in the publishers crawled. As shown in Table I, the snapshots in our datasets include different subsets of the top websites. Also, considering that the Alexa list is volatile and that it is constantly changing [34], we focus our analysis on a subset of 5100 publishers that are common across all time snapshots. **HTTPS Adoption**. Before digging deeper into the patterns and the effects of the online tracking ecosystem, here we make a preliminary measurement of the HTTPS adoption by the publishers. In our first snapshot (i.e., September 2017) we found that 3144 (61.64%) of the publishers were using HTTPS. This percentage is reasonable considering that our corpus was formed by publishers that rank among the Top-10K of the Alexa list, and that those tend to be more regulation compliant and probably privacy-preserving, as also reported in previous work [35]. In the snapshot of January 2018, we found 944 additional publishers (80.15% of our dataset) that adopted HTTPS, while ∼92% of the publishers adopted it by April 2019, validating the shift towards more secure practices. **Tracking Activity**. Figure 1 summarizes different dimensions of the tracking activity of 3rd-parties that are present in the

publishers of our datasets. The unique number of trackers embedded in each publisher (i.e., based on the outgoing HTTP requests observed) is shown in Figure 1(a). We observe that, in general, there is a clear decrease in the number of trackers through time. Specifically, until January 2018, 50% of the publishers communicated with almost 10 individual trackers in every visit, but this changed by the end of April 2019, when this number dropped to almost half, reaching the lowest measured value. This might not be a typical trend of the ecosystem, and this "momentary" elimination of trackers could be caused by the enforcement of GDPR in May 2018, but further investigation is needed to see whether publishers will go back to the same levels of tracking as before GDPR.

In Figure 1(b) we report the *total* number of 3rd-party requests per publisher. We observe that the trend on the distributions is similar: 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 show the *average* number of 3rd-party requests per publisher. In theory, since both the number of trackers and the 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, but for a small set of the publishers in the last snapshot (∼10%), this number increases.

Since 3rd-party requests may be issued for different reasons (e.g., cookie delivery, tracking pixels, content, ad-libraries), in Figure 1(d), we report the unique number of trackers that deliver 3rd-party libraries (i.e., JS) to publishers. For this measurement, we consider that connections to different subdomains, e.g., *subdomain1.domain.com/ad-library.js* and *subdomain2.domain.com/ad-library.js*, result in a connection to a unique 3rd-party provider. From the results, we note that by 2017, up to 50% of the publishers communicated with at least 5 different providers, and by 2019 this number decreased to 4. Also, there is a reduction in the plurality of the embedded providers (e.g., Amazon and Tapad were not present in the latest snapshots), which potentially results in higher tracking activity by the remaining entities, or increases the chances of cookie sharing and synchronization. Previous works on this topic (e.g., [14], [30], [36]) also report similar decline in the number and frequency of trackers. Specifically, in [36] 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 Web tracking ecosystem.

V. Publishers & Trackers: The PT Graph

Graph Construction. We follow a similar approach to Kalavri et al.'s [29] for constructing our graphs. We create a set of *2-mode* graphs of the publishers and their 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.

We represent all the domains that a browser requests as a 2-mode graph, by creating a set of mappings. The V_P represents the set of websites (publishers) a user visits, and accordingly V_T is the set of trackers embedded in publishers. We also define E_w as the set of weighted edges connecting vertices of the two different modes, and $w = (i, j)$ as the weight of the edge connecting tracker i with publisher j. Furthermore, we go beyond the state-of-art (i.e., [29]) and add weights on the edges, to represent the number of HTTP(s) requests between a publisher and a tracker. 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 reported in Section III, in each snapshot, we crawled a subset of the top Alexa list. Since we want to create a connected representation of the bipartite graphs, we use the Largest Connected Component (LCC) of each graph. We found that there are some isolated groups of nodes, that include websites that communicated with one or two different, but not popular trackers; we exclude such isolated groups from the graphs. In each of our final graphs, the LCCs contain on average ∼99% of the publishers and ∼95% of trackers of the originally crawled lists.

This type of connected graphs allows us to apply various graph metrics to quantify their properties, and compare them across time. In our analysis, we use graph metrics similar to those of previous works on this topic (e.g., [7], [29], [30]). Specifically we compute the Density and Diameter to study the inner structure of each network and the connectivity between their nodes. Other metrics like the Average Clustering Coefficient, Degree Centrality and Betweenness Centrality reveal the properties of each node in terms of the shortest paths that pass through the specific edge, and the number of connected neighbors. Finally, we also use Coreness-Periphery, a metric not studied before in the literature, that measures the importance of each node in terms of how "involved or core" it is inside the network. Detailed definitions of these metrics can be found in [37]–[39]. The generated graph models and their properties for each snapshots are given in Table II.

A. Stability of PT graph properties over time

The differences in the number of nodes and edges between these graphs reflect the different number of visited publishers in each snapshot. In general, the average clustering coefficient measures the degree with which the nodes of a graph tend to cluster together (i.e., tend to close triangles between triplets of

TABLE II: PT graph characteristics: number of vertices (N); edges (E); normalized average weight per edge (W); average clustering coefficient (CC); density (DE); diameter (D).

Dataset	N	E	W	CC	DE	D
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

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 [40]. This metric, in conjunction with the low density, reveals sparse connections between the different groups of nodes.

Regarding the number of trackers, there is a stable reduction through time, matching the measured decrease in the average number of trackers, as discussed in Section IV. In general, the characteristics and 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 reaches its lowest value, while the number of edges increases. This trend captures an increase in the connectivity of the graph nodes, pointing to tracker nodes being closer in the graph, as also discussed next.

Stability of Publishers vs. Trackers. We measure the degree, 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. Figure 2, plots the distribution for each metric, as computed on each snapshot for publishers and trackers.

Focusing our analysis on Figures 2a and 2d, the degree centrality for publishers is an order of magnitude lower than the one for trackers. Also, 50% of the publishers' degree centrality is ≤ 0.025, with only ∼ 5% of the publishers having more than 0.01. Conversely, tracker nodes have higher degree centrality scores, with ∼ 5% measured at ≥0.15, hinting to the fact of the well-known trackers that cover approximately all the publishers (e.g., Google and Facebook). Some examples of publishers with high 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 sets of nodes with different sizes and edge weights, it is reasonable for the tracking nodes to be more central in the network structure.

The betweenness centrality for publishers and trackers are given in Figures 2b and 2e, respectively. Publishers' low scores are expected, as betweenness centrality measures the extent to which a node lies on paths between other nodes. Obviously, publishers are not connected to each other but only with trackers. Trackers that are highly active, central and well known, are found on the tail of the distribution,
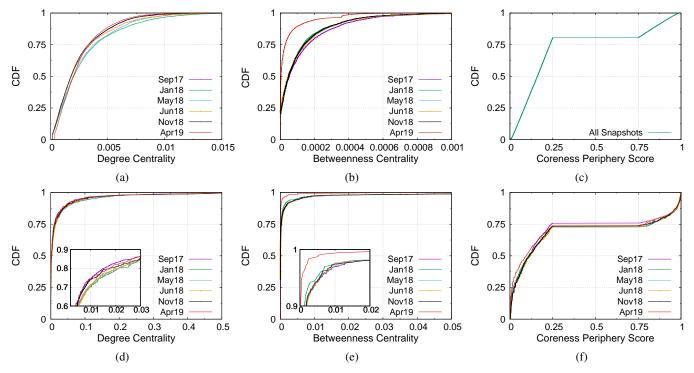
Fig. 2: Top row shows metrics for Publishers and bottom row for Trackers: (a) and (d) Normalized degree centrality; (b) and (e) Betweenness Centrality; (c) and (f) Coreness periphery for each PT Graph.

with scores ≥ 0.02. Example of publishers with high betweenness are *starbucks.com*, *livescore.com* and *cnn.com*, and accordingly trackers with high betweenness are *moatads.com*, *instagram.com* and *scorecardresearch.com*.

Finally, when comparing coreness scores for publishers and trackers, in Figures 2c and 2f, respectively, trackers tend to occupy positions in the graph with higher coreness than publishers. Interestingly, we observe 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 as *sfgate.com*, *sport.es* and *indianexpress.com*, and trackers such as *yandex.ru*, *adroll.com* and *amazon-adsystem.com*. have high coreness-periphery scores.

B. 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.

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 III. In general, we find a strong association between all the distributions with high confidence

TABLE III: Pearson Correlation Coefficient between the distribution of: Degree & 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. All results reported are statistically significant at $p-value<0.009$. The publishers are ranked according to the Alexa list, while trackers according to their Degree Centrality.

Dataset	Publishers			Trackers		
	$\sim DC\text{-}BC$	$\sim DC\text{-}CP$	$\sim BC\text{-}CP$	$\sim DC\text{-}BC$	$\sim DC\text{-}CP$	$\sim BC\text{-}CP$
Sep17	0.50	0.69	0.26	0.75	0.47	0.26
Jan18	0.43	0.83	0.3	0.73	0.50	0.24
May18	0.47	0.75	0.20	0.73	0.50	0.24
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

level. Trackers have higher correlation score in the Degree-Betweenness comparison, hinting their importance and central role on the network, regardless of the 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 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.

TABLE IV: 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. (*): Trackers who were Top-25 trackers across all snapshots. (+/-): Trackers who were 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

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

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

Furthermore, in Table V, we make a similar investigation for trackers' betweenness centrality. We note that the top of the list is populated by similar trackers. However, new entities such as *linkedin.com*, *ads-twitter.com* and *everesttech.net* emerge, demonstrating 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 and GDPR-compliant trackers.

VI. TRACKER TO TRACKER: THE TT GRAPH

Graph Construction. We build tracker-to-tracker graphs that are undirected but weighted, $TT = (V_{TT}, E_{TT})$, and originate from their corresponding PT graphs. Similarly to Section V, we create different sets of nodes that have specific properties. In our TT graphs, V_{TT} represents the set of trackers embedded in publishers, and E_{TT} is the set of weighted edges connecting two trackers, if and only if both trackers coexist in at least two different publishers. Also, the weight $w=(i,j)$, encodes the number of publishers that tracker i and j coexisted. A detailed description of our TT graphs is presented in Table VI.

C. What is the ecosystem's current state?

After analyzing the graphs' inner properties, we want to quantify the importance of the top tracking nodes for the ecosystem and their impact to user privacy. To have a clear view of the most important trackers at each time of our crawl, we measured their degree centrality and ranked them accordingly. A complete report about the Top-25 trackers across time is given in Table IV. 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 well established such as *Bluekai* and *Taboola*. This table also reports those trackers that gradually climbed into the Top list, illustrating the plurality of the tracking ecosystem.

TABLE VI: 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

TABLE VII: Percentage of overlap between the different sets of trackers for the CS edges extracted from [10]. 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

The TT graph of each snapshot has a fairly dense structure (average Density $0.19 - 0.21$). 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 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 3rd-parties potentially share user's information.

A. Are cookie synchronization pairs present in TT graphs?

Since the purpose of the TT graphs is to study potential data sharing among trackers, we need to compare the constructed pairs with existing data that already measure data sharing flows (i.e., ground truth datasets). We received access to two CS datasets provided by Papadopoulos et. al. [10] and Bashir et al. [7]. These datasets contain pairs of 3rd-parties performing CS while real users [10], or crawlers [7] browsed the web.

Interestingly, the dataset from [10] 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 [10], i.e., $w(i,j)$ is the number of times that a pair of trackers (i, j) performed information exchange. The first CS graph from [10] has 4656 trackers and 8582 edges connecting them, whereas the second CS graph from [7] has 59 trackers and 200 edges connecting them. To investigate the existence of CS pairs into TT pairs, we define the following sets:

- E_{CS} for the set of edges in a CS graph; E_{TT} for the set of edges in a TT graph.
- $\neg CS$ for the set of non-edges in a CS graph; $\neg TT$ the set of non-edges in a TT graph.

We also define the following overlaps of the above sets: $O_{common} = E_{CS} \cap E_{TT}, O_{\neg CS} = \neg CS \cap E_{TT}$ and $O_{\neg TT} = E_{CS} \cap \neg TT$. To have an accurate measurement between the different overlaps, we filter the edges and store only those that are part of the common trackers between each CS and each TT graph. A detailed report on the percentages of overlap between each of the two CS graphs is given in Tables VII and VIII.

TABLE VIII: Percentage of overlap between the different sets of trackers for the CS edges extracted from [7]. 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

According to Table VII, the overlap between CS and TT edges across snapshots is $47 - 60\%$. In the smaller CS dataset (Table VIII), this overlap is even higher, ranging to $64 - 81\%$. We remind the reader that this overlap is very high, considering all possible ways that such a graph could have been wired just by chance. Considering that the TT graphs were built artificially, 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 is a crucial finding: we can detect potentially collaborating pairs of trackers, without the need to deploy infrastructure to collect real users' data, or train artificial personas to collect CS activity. Also, the overlap is reduced over time possibly because of two reasons. First, the ground truth datasets were collected closer to our initial crawl (2017). Second, as reported by [30], there has been a reduction in CSs after GDPR. Therefore, the CS ecosystem is changing, and 3rd-parties that engaged in CS in earlier snapshots may not be doing so later, affecting the measured pair overlap.s Moreover, since the first CS graph was weighted, we investigated how well TT edges that overlap with CS edges cover the weights distribution. That is, how representative TT edges are of CS edges, with respect to weights. We found that the common TT edges cover well ($\sim$75%) of the distribution of the CS weights. Finally, 2% of the overlapping TT edges are the 5% most frequent edges on the CS graph, hinting again that synthetic data can capture confirmed real-world CS cases.

VII. Discussion & Limitations

We presented a first of its kind longitudinal study on the changes of the web tracking ecosystem during the last 2-3 years. The analysis through time was performed using six crawls of the top Alexa websites, with snapshots collected a few months apart. In the first level of analysis, we focused on network-level traffic between publishers and trackers. We found that there are fewer trackers embedded in the websites through time, with a reduction of 9% for the median site, and 10% for the 90% percentile site. Moreover, there are fewer 3rd-party HTTP(s) requests, with a reduction of 17% for the median and 25% for the 90% percentile site.

In the second level of analysis, we constructed bipartite graphs of publishers connected with their trackers (PT) and studied the PT graph properties through time. In summary, 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. Also, top trackers in terms of publisher coverage and centrality in PT graphs (e.g., *google-analytics*, *doubleclick*, *facebook*, *criteo*, *appnexus*, etc.), remained top through time. In terms of node importance, tracker sets appeared to be central in many of the examined centrality metrics.

Finally, we constructed tracker-to-tracker graphs (TT) for trackers who coexisted in the same websites. We compared these models with confirmed cookie synchronization (CS) pairs, which we use as ground truth flows between tracking entities. We found high overlap between the TT edges and the CS pairs, which means that we can detect potential CS activity and data sharing flows between trackers with a practical and cheaper alternative than collecting data from real users. Furthermore, the investigation of TT graph properties uncovered "hidden", and not studied so far, relationships between CS and TT pairs.

Limitations. The measurements presented in this paper constitute initial steps, along with related works, towards exploring the evolution of online tracking and the ad-ecosystem. Measuring such changes and identifying the tracking patterns and behaviors of the Web is a non-trivial process, as it can involve various vantage points and types of network connections, devices, user demographics, emulated and/or real users, etc. We believe that our study provides a lower bound on the intensity of the tracking ecosystem's state, but more work is needed towards establishing a full and representative picture of the online tracking, and its changes through time. Thus, our work can be extended in the following directions:

- **Main Page Crawls**. We performed multiple, consistent, easily replicable and lightweight- in terms of computing power and network usage- crawls on each website's main page. Indeed different tracking may be happening in subdomains and in nested website's layers, compared to the main domain page. Although it is important to go beyond the main page, it is also challenging, as it requires more resources and careful planning on browsing behav-

ior and activity exhibited by the crawlers (e.g., random vs. consistent, visit duration, etc.). Interestingly, this aspect of the problem was covered by recent work of Urban et. al. [41], where they measured the 3rd-party tracking mechanisms in subdomains of 10k popular pages. However, they did not investigated the problem in a time-longitudinal fashion. Thus, future work could study how subdomains' tracking ecosystem evolves through time.

- **IP Address & Location**. We measured the ecosystem's tracking state using an EU-based IP address, in order to highlight the potential GDPR effect on the EU space. Measuring biases, discrimination and other differences in tracking between locations inside and outside the EU (e.g., EU vs. USA, or EU vs. Asia) and the type of IP address used (e.g., residential, cellular, etc.) is a complex research problem in itself with several parameters at play, and it was out of the scope of this paper. In fact, recent works (e.g., [16], [42]) have investigated this problem in different ways, but not in a time-longitudinal fashion. Therefore, future work could study this tracker evolution while changing the crawling location.

- **Cookie Synchronization Detection**. One could argue that our crawling method is not able to capture the full scale of the cookie synchronization ecosystem. Indeed, measuring this kind of information exchange is challenging and requires specialized methodology such as using user synthetic personas [22], [43]. Instead, our approach is simpler and relies on automated crawls without personas, which captures much of the cookie synchronization, but not all. Thus, it measures a lower bound of such data flows, as would be expected for real users [10]. Future work could focus on utilizing more elaborated techniques on top of ours, to more accurately capture and study Cookie Synchronization flows.

VIII. Conclusion

In this paper, we presented a time-longitudinal study of the online tracking ecosystem, in order to analyze its changing state over a period of two years. We revealed patterns of the Web tracking, and effects of privacy laws such as GDPR on the tracking's intensity in over 5000 popular websites. Apart from studying such trends, we also presented graph models easily built over our automated crawls. These graphs capture different aspects of the tracking ecosystem, including key or central trackers with high embeddedness with respect to connectivity with other trackers, their influence on websites, as well as changes in their ranking of importance. We also studied how information flows between trackers, due to cookie synchronization in traffic of real users, are reflected and match flows found in our automated (non-real user) crawls. Our study points to the future opportunity for building accurate and representative graph models of the Web tracking ecosystem, without the overhead of collecting real user data.

REFERENCES

[1] F. Roesner, T. Kohno, and D. Wetherall, "Detecting and defending against third-party tracking on the web," in *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 12–12. [Online]. Available: http://dl.acm.org/citation.cfm?id=2228298.2228315

[2] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 1388–1401. [Online]. Available: http://doi.acm.org/10.1145/2976749.2978313

[3] P. Eckersley, "How unique is your web browser?" in *Proceedings of the 10th International Conference on Privacy Enhancing Technologies*, ser. PETS'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 1–18. [Online]. Available: http://dl.acm.org/citation.cfm?id=1881151.1881152

[4] G. Acar, M. Juarez, N. Nikiforakis, C. Diaz, S. Gürses, F. Piessens, and B. Preneel, "Fpdetective: Dusting the web for fingerprinters," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, ser. CCS '13. New York, NY, USA: ACM, 2013, pp. 1129–1140. [Online]. Available: http://doi.acm.org/10.1145/2508859.2516674

[5] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "Cookieless monster: Exploring the ecosystem of web-based device fingerprinting," in *Proceedings of the 2013 IEEE Symposium on Security and Privacy*, ser. SP '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 541–555. [Online]. Available: http://dx.doi.org/10.1109/SP.2013.43

[6] L. Olejnik, T. Minh-Dung, and C. Castelluccia, "Selling off privacy at auction," in *Network and Distributed System Security Symposium (NDSS)*, 2014.

[7] M. A. Bashir and C. Wilson, "Diffusion of user tracking data in the online advertising ecosystem," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 4, pp. 85–103, 2018.

[8] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The web never forgets: Persistent tracking mechanisms in the wild," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '14. New York, NY, USA: ACM, 2014, pp. 674–689. [Online]. Available: http://doi.acm.org/10.1145/2660267.2660347

[9] P. Papadopoulos, N. Kourtellis, and E. P. Markatos, "The cost of digital advertisement: Comparing user and advertiser views," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18. International World Wide Web Conferences Steering Committee, 2018, p. 1479–1489.

[10] P. Papadopoulos, N. Kourtellis, and E. Markatos, "Cookie synchronization: Everything you always wanted to know but were afraid to ask," in *The World Wide Web Conference*. ACM, 2019, pp. 1432–1442.

[11] Council of the European Union , European Parliament, "Directive 2002/58/ec of the european parliament and of the council of 12 july 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (directive on privacy and electronic communications)."

[12] ——, "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.

[13] "California consumer privacy act of 2018 [1798.100 - 1798.199]," 2018.

[14] C. Iordanou, G. Smaragdakis, I. Poese, and N. Laoutaris, "Tracing cross border web tracking," in *Proceedings of the Internet Measurement Conference 2018*. ACM, 2018, pp. 329–342.

[15] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz, "We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy," *arXiv preprint arXiv:1808.05096*, 2018.

[16] A. Dabrowski, G. Merzdovnik, J. Ullrich, G. Sendera, and E. Weippl, "Measuring cookies and web privacy in a post-gdpr world," in *International Conference on Passive and Active Network Measurement*. Springer, 2019, pp. 258–270.

[17] J. Sørensen and S. 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: Association for Computing Machinery, 1 2019.

[18] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz, "(un) informed consent: Studying gdpr consent notices in the field," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 973–990.

[19] P. Agarwal, S. Joglekar, P. Papadopoulos, N. Sastry, and N. Kourtellis, "Stop tracking me bro! differential tracking of user demographics on hyper-partisan websites," in *Proceedings of The Web Conference 2020*, ser. WWW '20, 2020, p. 1479–1490.

[20] S. Zimmeck, J. S. Li, H. Kim, S. M. Bellovin, and T. Jebara, "A privacy analysis of cross-device tracking," in *26th USENIX Security Symposium*, ser. USENIX Security 17. USENIX Association, 2017, pp. 1391–1408.

[21] J. Brookman, P. Rouge, A. Alva, and C. Yeung, "Cross-device tracking: Measurement and disclosures," *Proceedings on Privacy Enhancing Technologies*, 2017.

[22] K. Solomos, P. Ilia, S. Ioannidis, and N. Kourtellis, "Talos: An automated framework for cross-device tracking detection," in *22nd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2019)*, 2019.

[23] J. R. Mayer and J. C. Mitchell, "Third-party web tracking: Policy and technology," in *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, ser. SP '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 413–427. [Online]. Available: http://dx.doi.org/10.1109/SP.2012.47

[24] M. Falahrastegar, H. Haddadi, S. Uhlig, and R. Mortier, "Tracking personal identifiers across the web," in *International Conference on Passive and Active Network Measurement*. Springer, 2016, pp. 30–41.

[25] Z. Yu, S. Macbeth, K. Modi, and J. M. Pujol, "Tracking the trackers," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016, pp. 121–132. [Online]. Available: https://doi.org/10.1145/2872427.2883028

[26] N. Nikiforakis, W. Joosen, and B. Livshits, "Privaricator: Deceiving fingerprinters with little white lies," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2015, pp. 820–830. [Online]. Available: https://doi.org/10.1145/2736277.2741090

[27] A. Panchenko, F. Lanze, J. Pennekamp, T. Engel, A. Zinnen, M. Henze, and K. Wehrle, "Website fingerprinting at internet scale." in *NDSS*, 2016.

[28] A. Lerner, A. K. 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 *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, 2016. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/lerner

[29] V. Kalavri, J. Blackburn, M. Varvello, and K. Papagiannaki, "Like a pack of wolves: Community structure of web trackers," in *International Conference on Passive and Active Network Measurement*. Springer, 2016, pp. 42–54.

[30] T. Urban, D. Tatang, M. Degeling, T. Holz, and N. Pohlmann, "The unwanted sharing economy: An analysis of cookie syncing and user transparency under gdpr," *arXiv preprint arXiv:1811.08660*, 2018.

[31] I. Sanchez-Rola, M. Dell'Amico, P. Kotzias, D. Balzarotti, L. Bilge, P.-A. Vervier, and I. Santos, "Can i opt out yet? gdpr and the global illusion of cookie control," in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 2019, pp. 340–351.

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

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

[34] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez, "A long way to the top: significance, structure, and stability of internet top lists," in *Proceedings of the Internet Measurement Conference 2018*. ACM, 2018, pp. 478–493.

[35] A. P. Felt, R. Barnes, A. King, C. Palmer, C. Bentzel, and P. Tabriz, "Measuring https adoption on the web," in *USENIX Security*, 2017.

[36] J. Sørensen and S. Kosta, "Before and after gdpr: The changes in third party presence at public and private european websites," in *The World Wide Web Conference*. ACM, 2019, pp. 1590–1600.

[37] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[38] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[39] P. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha, "Core-periphery structure in networks (revisited)," *SIAM Review*, vol. 59, no. 3, pp. 619–646, 2017.

[40] M. Latapy, C. Magnien, and N. Del Vecchio, "Basic notions for the analysis of large affiliation networks/bipartite graphs," *arXiv preprint cond-mat/0611631*, 2006.

[41] T. Urban, M. Degeling, T. Holz, and N. Pohlmann, "Beyond the front page: Measuring third party dynamics in the field," in *Proceedings of The Web Conference 2020*, 2020, pp. 1275–1286.

[42] D. Zeber, S. Bird, C. Oliveira, W. Rudametkin, I. Segall, F. Wollsén, and M. Lopatka, "The representativeness of automated web crawls as a surrogate for human browsing," in *Proceedings of The Web Conference 2020*, 2020, pp. 167–178.

[43] J. M. Carrascosa, J. Mikians, R. Cuevas, V. Erramilli, and N. Laoutaris, "I always feel like somebody's watching me: measuring online behavioural advertising," in *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies (CONEXT)*, 2015.