Measuring the Impact of the GDPR on Data Sharing in Ad Networks

Tobias Urban
urban@internet-sicherheit.de
Institute for Internet Security
Ruhr University Bochum

Dennis Tatang
dennis.tatang@rub.de
Ruhr University Bochum

Martin Degeling
martin.degeling@rub.de
Ruhr University Bochum

Thorsten Holz
thorsten.holz@rub.de
Ruhr University Bochum

Norbert Pohlmann
pohlmann@internet-sicherheit.de
Institute for Internet Security

ABSTRACT

The European General Data Protection Regulation (GDPR), which went into effect in May 2018, brought new rules for the processing of personal data that affect many business models, including online advertising. The regulation's definition of personal data applies to every company that collects data from European Internet users. This includes tracking services that, until then, argued that they were collecting anonymous information and data protection requirements would not apply to their businesses.

Previous studies have analyzed the impact of the GDPR on the prevalence of online tracking, with mixed results. In this paper, we go beyond the analysis of the number of third parties and focus on the underlying information sharing networks between online advertising companies in terms of client-side cookie syncing. Using graph analysis, our measurement shows that the number of ID syncing connections decreased by around 40 % around the time the GDPR went into effect, but a long-term analysis shows a slight rebound since then. While we can show a decrease in information sharing between third parties, which is likely related to the legislation, the data also shows that the amount of tracking, as well as the general structure of cooperation, was not affected. Consolidation in the ecosystem led to a more centralized infrastructure that might actually have negative effects on user privacy, as fewer companies perform tracking on more sites.

CCS CONCEPTS

• **Security and privacy** → *Privacy protections*; *Privacy protections*;
• **Social and professional topics** → *Privacy policies.*

KEYWORDS

cookie syncing; GDPR; privacy; online advertisement; tracking

ASIA CCS '20, June 1–5, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-6750-9/20/06…$15.00
https://doi.org/10.1145/3320269.3372194

ACM Reference Format:
Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Measuring the Impact of the GDPR on Data Sharing in Ad Networks. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security (ASIA CCS '20), June 1–5, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3320269.3372194

1 INTRODUCTION

Advertising remains one of the main sources of income for many websites, apps, and online services. Many business models rely on ads and analytics services [44] to personalize their products and to be able to offer them "for free". To individually target website visitors with ads, tracking services gather personal data, mostly without users' explicit consent [51]. Personalized ads are based on data collected by ad companies about Internet users through various mechanisms, mainly HTTP cookies [1, 16]. The gathered data is often seen as an economic asset of a company [42]. But attackers also perform malicious exfiltration of personal data [54]. As a result, the imbalance of power between data processors (service providers) and data subjects (users) increased in the last couple of years. Users are often not aware of the collection, usage, or consequences of the use of their data [11] and have only limited options when trying to control it [41]. To address some of these problems, the European General Data Protection Regulation (GDPR), which went into effect on May 25, 2018, introduced significant changes that affect how personal data can be collected and shared. Compliance with the GDPR rules is required for any company that offers services in the European Union—no matter where their headquarters are located [17].

In this work, we seek to provide insights into the effects of the GDPR on the information sharing behavior between ad services. Previous studies have described how cookie syncing is used to share identifiers [1, 16], but there is a lack of knowledge about its extent, the networks behind it, and its development over time. More specifically, we measure the relations of websites and third parties, as well as links between third parties regarding ID syncing before and after the GDPR took effect. Over the course of our experiment, we used different browser profiles to visit more than 2.6 million websites (⌀ 221,000 in each crawl; 8,000 unique domains) over the course of ten months to identify ID syncing between third parties embedded in these websites. We use graph analysis techniques to measure connections between third parties with respect to ID syncing and

demonstrate a decrease in the number of sharing communities and the betweenness centrality, a measure for information flows.

Recent work has found that around the date of GDPR enforcement in May 2018, the adoption of privacy policies and cookie notices increased [15] and that at the same time the amount of tracking [8] and cookie usage [12] decreased. However, others found that the effects of the GDPR are not as significant in terms of directly embedded third parties [43]. Our analysis shows that changes can be observed if more complex coherences are taken into account rather than counting third parties. We perform an in-depth analysis to measure the effects of the new legislation on the tracking ecosystem as we investigate links between companies and go beyond the measurements that focus on embedded third parties and cookies directly set by websites. We show that while the amount of data collected about Internet users may not have changed since May 2018, the number of online advertising companies that share information has decreased. At the same time, those that still share information have not limited their efforts, instead, some companies might benefit from an ongoing centralization.

To summarize, our study makes the following contributions:

- We measure changes regarding the use of third-party services by websites shortly before and the months after the GDPR enforcement and show the shift of relations between these third parties in terms of ID syncing. Based on twelve measurements over a period of ten months, starting before the GDPR's enforcement date, we show that the amount of links between companies is reduced by over 40 %.
- We employ methods of graph analysis to construct an undirected graph that describes the relations between third parties. Different measures of ID sharing communities show that the general structure of relations is not affected.
- Finally, we analyze the topology how third parties are connected and show that third parties are often arranged in star-like topologies with one central node that is sometimes linked to hundreds of outer nodes.

The remainder of the paper is organized as follows: We first give an overview of online tracking, cookie syncing, and the GDPR (Section 2) and then discuss how our measurements compare to related work (Section. 3). Afterwards, we describe our measurement framework using `OpenWPM` (Section 4). Our results section (Section. 5) describes the changes in the advertising ecosystem we observed and offers some explanations for them (Section 6). We discuss potential limitations of our analysis and conclude with a summary of our results (Section 6).

2 BACKGROUND

In this section, we provide some background information necessary to study the effects of legislation on information sharing between online advertising companies. We describe the overall advertising ecosystem, technical details of cookie syncing, and the importance of the GDPR for this socio-technical system.

2.1 Advertising Economy

Displaying ads is the most common way to fund online services. in 2017, the online advertising industry generated total revenues of $88.0 billion [26] in the US and €41.8 billion in the European Union [25]. The ecosystem behind this is complex and consists in a nutshell of three basic entities described in the following [59].

On the one end, there are publishers and website owners that use *supply-side platforms* (SSP) to sell ad space on websites. On the other end, the *demand-side platform* (DSP) is used by marketing companies to organize advertising campaigns, across a range of publishers. To do so, they do not necessarily have to select a specific publisher they want to work with but can define target users based on different criteria (e. g., geolocation, categories of websites visited, or personal preferences). A *data management platform* (DMP) captures and evaluates user data to organize and optimize digital ad campaigns. They can be used to merge data sets and user information from different sources to automate campaigns on DSPs. To do so, a DMP often collects IDs of different systems and merges data with those from other sources to target ad campaigns to a specific audience based on high-level information like interest profiles [14].

Therefore, user tracking and profiling are critical parts of website and mobile application business models alike [1, 16, 44]. Profiles containing information necessary to target advertisements like interests or lists of previous purchases are often based on the users' clickstream (a list of websites a user has visited) to enable targeted advertising [7]. A unique digital identifier is assigned to each user, either by a server or computed based on properties of the user's device (*device fingerprinting* [16]). The most prevalent way to store such digital identifiers on a user's device are *HTTP cookies*.

2.2 Cookie Syncing

A *HTTP cookie* is a piece of textual data, strictly limited in size, that can be set by a website to store data locally on a client. In theory, cookies contain simple `name=value` pairs but in practice, they often serve as a reference (i. e., a user ID) and combine information through various means [21]. Cookies are intended to maintain a state between different HTTP sessions, e. g., to remember user preferences, to keep items stored in the shopping cart, or to log that a user has previously authenticated with the server. Storing a unique user identifier in a cookie allows a server to identify a user revisiting a website. It is also common that additional information exceeding the allowed size for cookies is stored on the server related to that same ID (e. g., inferred interest segments). If the website originally opened by a user sets a cookie, it is called a *first-party cookie* (`A` in Figure 1). A cookie is called a *third-party cookie* if the visited website embeds an object from another domain and this third party sets a cookie (`B1` and `B2` in Figure 1). For online advertising, this could be profile information like inferred interest segments or geolocation. A server can only access a cookie under the domain that set it, meaning that different third parties cannot access each other's cookies. This prohibits data leakage or cross-domain tracking of different third parties by merely accessing the cookies (via the *Same-Origin Policy*).

Cookie syncing is a process to bypass the Same-Origin Policy by sharing the unique identifier of a user between two third parties (`C` in Figure 1). Cookie syncing is mostly a two-step process: (`C1`) a script from a third-party (`bar.org`) is loaded into a website (`example.org`). (`C2`) The request that loads the script is then redirected or the script itself issues a new request to the syncing partner

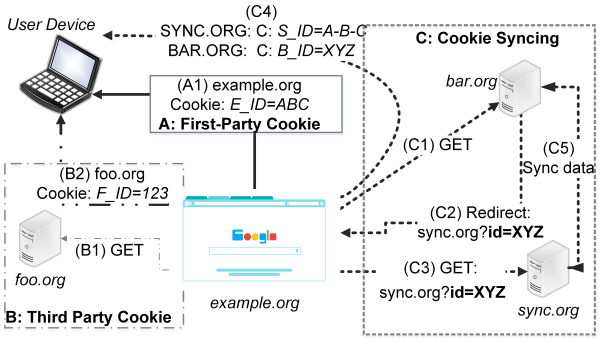


Figure 1: Different types of cookies: (A) a first-party cookie—directly set by the visited website, (B) a third-party cookie—set by a third party embedded in the website, and (C) a synchronized cookie—shared between two parties.

(`sync.org`). This redirected request contains the ID `bar.org` assigned to the user (e. g., `sync.org?bar_user_id=XYZ`). After this ID syncing `sync.org` knows, via the HTTP referrer header or additional information added to the request, that the user with `bar.org`'s ID visited `example.org` (C3). If `sync.org` already has a cookie (e. g., from a previous visit to another website) on the client, it can map `bar.org`'s user ID with its own (C4). This allows `sync.org` and `bar.org` to share data about the user over another channel (C5). This mechanism also allows a tracking company (`sync.org`) to track users on a large variety of websites even if these websites do not directly embed a tracker by that company but by its partners.

While this is considered an undesirable privacy intrusive behavior by some, it is in practice a fundamental part of the online ad economy to perform *Real-Time Bidding* (RTB) [36]. In RTB impressions and online ad space are sold in real-time on automated online marketplaces whenever a website is loaded in a browser.

2.3 Legal background

The General Data Protection Regulation (GDPR or Regulation 2016/679) [17] is a regulatory instrument by the European Union (EU) to harmonize data protection laws between its member states. After a transition period of two years, it was put into effect on May 25, 2018. The GDPR specifies under which circumstances personal data may be processed, lists rights of data subjects, and obligations for those processing personal data of EU citizens. It is therefore important for all companies offering services which collect and process personal data in Europe. The GDPR was expected to have a strong impact on the online advertising ecosystem as it provides a broader understanding of what is considered to be personal data [37]. Until before the GDPR, many advertising companies claimed (and still claim) that they only process anonymized data because the profiles they use for targeted advertising mostly do not contain personal identifiers like names or home addresses. In contrast, GDPR considers this pseudonymous data as it still describes one single person that is re-identifiable with additional information.

The European Data Protection Authorities (Article 29 Working Group) had already decided in 2010 that profiles created through online tracking are considered personal data and would need explicit consent [13], but studies on Web tracking showed that online advertisers did not follow these recommendations, for example by ignoring the Do-Not-Track signal [48]. It was expected that the GDPR led to changes and influenced the online advertisement ecosystem since it extended its legal reach to companies that conduct business with the EU regardless of where their headquarters are located. Compared to previous legislation, it also allows data protection authorities to fine companies much higher than before with up to 4 % of their global annual revenue. In January 2019, the French data protection authority (CNIL) fined Google for 50 million Euros for not validly obtaining consent [10].

3 RELATED WORK

Multiple research groups have studied how websites and third-party tracking changed around the enforcement date of the GDPR in May 2018. In this section, we provide an overview of the related work on GDPR measurements and similar research in this area.

3.1 GDPR Measurements

An overview of privacy-related measurement studies, with a focus on the GDPR, is given in Table 1. Different studies measured similar topics with mixed results [8, 43]. This, alongside our results, highlights that effects of complex legislation (i. e., the GDPR) are not necessarily measurable in all parts of a complex ecosystem such as online advertising or online tracking. Our work differs from the related work as we do an in-depth analysis of the ecosystem (i. e., connections of third parties) and do not limit our measurements to the embedded third parties.

Most recent works measured the effect of the GDPR regarding cookie usage and embedded third parties. Dabrowski et al. measure the effects of cookies set based on the location of a user and find that around 50 % more cookies are being set if the users come from outside the EU [12]. In contrast, Sørensen et al. found that the number of third parties slightly declined since the GDPR went into effect (which is in line with our findings) but they conclude that the GDPR is not necessarily responsible for that effect [43].

Regarding GDPR rights, Urban et al. have shown that performing *subject access requests* (SARs) can be a tedious and often unsuccessful process [53] while the data received by SARs is often not intuitive and not helpful [52].

3.2 Online Privacy Measurements

Most previous work analyzes online privacy through measurements, which have *all* been conducted prior to the GDPR. For example, Gonzales et al. presented a large-scale study on the use of HTTP cookies [21]. The authors analyzed more than 5.6 billion HTTP requests over a period of 2.5 months. They show that, in practice, cookies are much more sophisticated than simple `name=value` pairs and present an algorithm capable of inferring the format of a cookie with high recall and precision rates. In 2016, Englehardt and Narayanan published their work on measuring online tracking [16]. They introduce the open-source measurement tool `OpenWPM`, which they used to crawl and analyze the top one million websites on the Internet. They analyzed cookie-based and fingerprint-based tracking along with 13 other types of measurements. Papadopoulos

Table 1: Overview of privacy measurements conducted after the GDPR took effect. ✓ indicates a measurable effect of the GDPR, while ✗ indicates the opposite. ☆ The work relates to the EU cookie directive and not the GDPR.

Author	Venue	Scale (websites visited)	Technology	Focus	Main finding	Study time frame	GDPR had effect
Degeling et al. [15]	NDSS'19	6,759	proprietary	Privacy policies & cookie notice	Right before the GDPR took effect, companies updated their privacy policies; cookie notices lack usability.	01/18–06/18	✓
Dabrowski et al. [12]	PAM'19	100,000	headless Chrome	Cookie usage	Websites set 49 % less cookies if users from the EU visit them.	06/18	✓
Sorensen et al. [43]	WWW'19	1,250	`OpenWPM`	Third party usage	Effects of the GDPR on third-party usage is not clear.	02/18–09/18	✗
Sanchez-Rola et al. [41]	AsiaCCS'19	2,000	manual collection	Cookie usage and consent	The GDPR has global reach (e. g., cookie banners) but tracking is often still present even if opted out.	07/18	✗
Trevisan et al. [50]	PETS'19	35,000	`CookieCheck` & `WebPageTest`	Cookie usage	49 % of websites do *not* honor the cookie directive.	04/17	✗☆
Cliqz [8]	Blog post	2,000	proprietary	Online tracking	Large trackers (slightly) gain in coverage while shares of smaller trackers (clearly) decreas.	03/18–07/18	✓
Libert et al. [33]	Technical report	10,168	`webXray`	Third party usage	News websites use less social media content; cookie usage, without consent, is decreased by around 22 %.	04/18–06/18	✓
Our Study	AsiaCCS'20	6,527	`OpenWPM`	Cookie syncing	GDPR has a statistically significant impact on cookie syncing, which is reduced by around 40%.	05/18–03/19	✓

et al. performed a study on cookie syncing on a dataset the collected over the course of one year including browsing activity from 850 mobile devices [39]. According to their measurement, over 97 % of users are exposed to cookie syncing and an ID is shared with 3.5 companies on average. Karaj et al. monitored the online tracking landscape over a period of ten months using data provided by real users through a browser extension [30]. They try to illuminate effects of the GDPR on the online tracking business and argue that more transparency and accountability is needed since users struggle to keep control of their data.

3.3 ID Sharing

In addition to the studies referenced in Subsection 3.2, work has been conducted regarding ad networks. Falahrastegar et al. investigated the connections between third parties focusing on ID sharing [18]. They found that domains show more syncing activities when a user is logged out and group the sharing domains based on their content. Most recently, Bashir et al. introduced a so-called *inclusion graph* that models the diffusion of online tracking through Real-Time Bidding [4]. They show that 52 advertisers or analytics companies obverse over 90 % of an average user's online clickstream. The work differs from ours since we do not want to shed light on the connection of online advertising companies but measure effects of the GDPR. A method to identify server-side information flow in the ad economy was presented by Bashir et al. [3]. They use re-targeted ads to reveal information flows.

3.4 Computer Law and Privacy Policies

Aside from the presented more technical papers, our work is related to work that focuses on the legal aspects of the GDPR. Recently, Libert presented his work on an automated approach to auditing disclosure of third-party data collection in websites' privacy policies [32]. The work shows empirically that it is unmanageable for a person to read the privacy policies of the first and third parties. De Hert et al. [24] discuss the right to data portability from a computer law point of view. De Hert et al. give a systematic interpretation of the new rights and propose two approaches how to interpret the legal term "data provided" in the GDPR. The authors describe a minimal approach, where only data directly given to the controller (e. g., data entered into a form) can be seen as "provided." They also describe a broad approach which also labels data observed by the controller (e. g., browser fingerprints) as "provided." The authors propose to adopt the extensive approach.

3.5 Distinction from Previous Work

The introduced related work measures the tracking capabilities and other privacy implications of websites—some in relation to the GDPR. However, previous work related to the GDPR simply looked at the third parties present on websites and if their presence changed [8, 43], measured tracking techniques and their prevalence [1, 16], or analyzed cookie setting practices of third parties [15, 41]. In this work, we go deeper and provide insights in the connections of third parties as far as these are observable on the client. We focus on the amount of sharing connections, the typologies how companies are related to each other, and provide

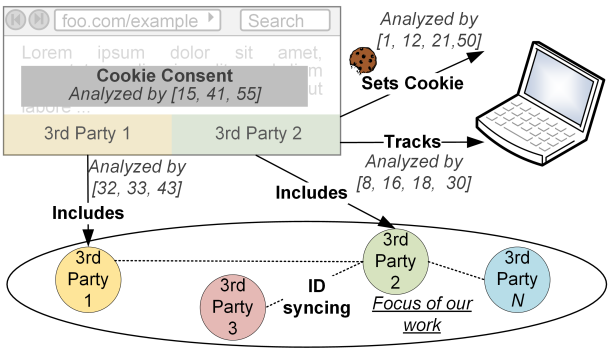


Figure 2: Overview of related work and how our work is distinct from it.

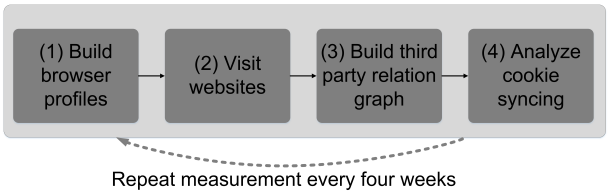


Figure 3: Overview of our measurement setup. First, we build the browser profiles which we use to visit the websites. Based on the captured traffic, we build the third-party graphs which we analyze regarding ID syncing.

some case studies on specific companies and how they adopted the new legislation. Figure 2 highlights our contribution and its distinction from previous work.

4 MEASUREMENT APPROACH

We conducted a measurement study of cookie syncing in the browser to gain insights into information sharing between tracking companies and the impact of the GDPR on these practices. In the following, we describe our measurement framework and explain how we measure the syncing relations of third parties.

4.1 Measurement Framework

To measure the extent of cookie synchronization and the existing networks in the sharing economy, we used the `OpenWPM` [16] platform. For our study, we deployed the platform on two computers at a European university to ensure a European origin of our generated web traffic. We chose not to use a scalable web service (e. g., Amazon EC2) to automate our measurement since it is easier for a website to detect such automated crawls [28]. Additionally, we conducted two additional measurements using US-based IP addresses using a VPN service to validate the effects of geolocation.

`OpenWPM` was configured to log all HTTP request and response headers, HTTP redirects, and POST request bodies as well as various types of cookies (e. g., Flash cookies). We did not set the "Do Not Track" HTTP header and allowed third-party cookies. We used simple bot detection mitigation techniques (i. e., scrolling randomly up and down on each visited website and randomly jiggling with the mouse) to make it more difficult to detect our crawler. As `OpenWPM` is an instrumentation of the Firefox browser, our measurement is limited to cookie syncing on the browser level.

In each subsequent measurement of our analysis, we created 400 browsing profiles. A "browser profile" is a separate browser instance with its own cookie store, caching, and browsing history. Each profile had its own browser storage to make sure cookies could be separately stored for each session. We created 20 profiles for the top 20 countries with the highest number of Internet users worldwide [27]. The top 20 countries account for 71 % of all Internet users. The list contains six countries from the EU, three countries from the Americas, six countries from Asia, and five countries from Africa and the Middle East. We choose to use the worldwide top countries, and not just EU top countries, since GDPR applies to all companies that offer services to EU residents. Furthermore, we randomly assigned a popular user-agent string and a common screen resolution[1] to each browser profile that remained constant during the crawling process per session. Each profile was assigned at random so that all 400 profiles used its own set of user agent and screen resolution (around 312 different combinations in each country). We used an artificially populated cookie store and browsing history in each browser profile which we created by browsing 100 random websites selected from the Alexa top 1000 list.

For each profile, we took the Alexa top 500 list of the corresponding country [2] (as off May 2018) and randomly chose 100 to 400 websites to be visited. We randomized the number of websites to mimic a more realistic user behavior and capture more realistic cookie syncing activities. During the course of all our measurements, we used the same Alexa top lists to allow better comparability across our measurements. We limited our measurement to the top 500 websites to be able to conduct measurements in a reasonable time (one measurement took about one week). In all measurements each website was visited with at least one profile and no websites excluded EU residents from their service (e. g., by showing error pages or sending HTTP error codes). To mimic interactions with the websites, we extracted all first-party links from their landing pages. For example, when visiting *foo.com*, we extracted all links to pages on *foo.com* and randomly visited two to four of those. In the remainder of this paper we call these links *subsites* since they are all associate with the same website but have a distinct URL. We decide to randomize the visited websites because we wanted to measure the effects of the new legislation on a broader scale and not just the effect of a chosen set of domains or sub sites. Overall, we visit between 120,000 and 800,000 (221,656 on average (SD 10,609)) distinct URLs per measurement. An overview of the measurement approach is given in Figure 3.

We conducted twelve measurements (M#1–M#12) over the course of ten months. The first measurement started just days before the GDPR went into effect (May 19, 2018), the second right after the GDPR went into effect (May, 25 2018). The following measurements were made in intervals of about four weeks (i. e., one measurement in the third calendar week (CW) of each month, from 05/18 to 03/19). We performed two reference measurements with US-based IP addresses via a VPN connection in October 2018 and January 2019

[1] User agents were collected from *TechBlog* [45], most common screen resolution set as reported by *Global Stats* counter [20].

to compare the results with Europe-based traffic from the same time. VPN services can potentially inject content (e. g., ads) into the traffic, which might affect the results [31]. However, the Terms of Service of the used VPN service (*NordVPN*) neither stated that this might happen nor did we find any information about content injection for this VPN service. To avoid dishonest statements of the VPN service provider regarding the location of their servers [57], we checked at the beginning of each experiment if the VPN service had assigned an IP address associated with an US geolocation using different services (e. g., "*IP Location Finder*" [29] or "*What Is My IP Address*" [58]) and monitored that this address did not change during the experiment. For each measurement, we use a newly created profile (i. e., new and different cookie stores) to avoid pollution of our dataset.

4.2 Identification and Mapping of Third-Party Relations

To analyze the sharing of *personal* or *digital identifiers* (IDs), we first need to define them. For every visited domain we analyzed the HTTP `GET` and `POST` requests and split the path or body of the requests at characters that are typically used as delimiters (e. g., '&' or ';'). As a result, we obtained a set of ID candidates we stored as key-value pairs for later analysis. We identified IDs according to the following rules inspired by Acar et al. [1]:

- Eliminate all ID candidates that were observed for multiple profiles. Every identifier should be unique to each profile (e. g., we eliminate $c1 = (p\_id, 1234abcd)$ and $c2 = (p\_id, 1234abcd)$ if they were observed in two profiles).
- Eliminate ID candidates with the same key but where values differ in length. We expected that IDs are of consistent length (e. g., the candidates $c1 = (data, 3rw3)$ and $c2 = (data, 70g63b5g)$ would be eliminated).
- Eliminate candidates whose values do not contain enough entropy (according to the Ratcliff/Obershelp pattern recognition algorithm [40]) to contain an ID. Since we only observe a small fraction of the potential ID space, we expect that IDs differ significantly (e. g., the candidates $c1 = (id, AAAC)$ and $c2 = (id, AABA)$ would be eliminated).
- Exclude candidates whose length is too short to contain enough entropy to hold an ID. To provide enough entropy, we expect an ID to have at least eight characters (e. g., the candidate $c = (key, 1hgtz)$ is excluded).

To measure the syncing relations of third parties, it is necessary to identify URLs in a request that contain user IDs (e. g., **foo.com**`/sync?partner=https://`**bar.com**`?/id=abcd-1234`). To do so we attempt to decode (e. g., BASE64) and deflate (e. g., `gzip`) every HTTP `GET` and `POST` argument. Since any of these arguments might be encoded/inflated multiple times, as observed by Starov et al. [44], we repeated this process multiple times (if necessary). We used regular expressions to parse the decoded values for URLs. When an URL was found, we check if this URL has GET parameters that might be an ID, according to our definition of an ID.

We used the *WhoTracks.me* database [9] to cluster all observed third-party websites based on the company owning the domain. These clusters served as nodes for the construction of an undirected graph. We added two types of edges to the graph to connect the nodes: (1) direct relations (i. e., a website embeds a third-party object) and (2) syncing relations (i. e., two third parties that perform cookie syncing). Thus, we can measure (1) how many websites make use of a specific third party and (2) with how many other third parties IDs were synced. If we found a request was used to sync user IDs, we created a link in the constructed graph for the measurement in which the syncing was observed.

5 RESULTS AND EVALUATION

To analyze the effects of the GDPR regarding cookie synchronisation, we performed monthly measurements between May 2018 and March 2019 (twelve in total). Excluding the US reference measurements, we visited 2,659,873 URLs in our study, resulting in over 1 TB of data, in terms of size of the `OpenWPM` databases. We refer to our first measurement as pre-GDPR measurement, because it was conducted before the GDPR went into effect, and to all other measurements as post-GDPR measurements. Based on the data gathered in our measurements, we created graphs to represent the ID sharing between different companies. The resulting graphs show a steep decrease in sharing after the GDPR went into effect.

Table 2 provides an overview of the size of each measurement, which varied due to some randomization introduced as described in Section 4. The table lists the number of domains visited in each measurement to allow for comparison of our results with related work. For the remainder of the paper, we cluster the observed third parties based on the respective owning owning company (see Section 4). Figure 4 illustrates the size of the pre-GDPR measurement in relation to the post-GDPR measurements. While the number of visited domains was above average (8,448) but within the interquartile range (25^{th} and 75^{th} percentile), the amount of actually visited websites in M#1 is above the median (but slightly below the average of 221,656) but also within the interquartile range.

In line with previous work [15, 43] our data shows that the average number of third parties embedded in websites did not change before and after the GDPR went into effect. But when considering the whole ecosystem, changes can be observed.

Table 2: Overview of our measurements. For each measurement the number of visited domains, the visited number of subsites, and the observed third parties are given.

ID	Date	CW	Domains	Subsites	⌀3rd P.
M#1	2018/05/19	20	8,576	220,948	5.22
M#2	2018/05/25	21	8,723	239,636	5.10
M#3	2018/06/18	26	8,073	204,108	5.17
M#4	2018/07/23	28	8,267	216,283	5.21
M#5	2018/08/20	34	8,278	212,405	5.22
M#6	2018/09/17	38	8,334	218,687	5.17
M#7	2018/10/22	43	8,629	225,230	5.23
M#8	2018/11/19	47	8,259	219,164	5.22
M#9	2018/12/21	51	8,680	223,718	5.27
M#10	2019/01/19	3	8,667	222,122	5.22
M#11	2019/02/18	7	8,424	215,407	5.26
M#12	2019/03/18	11	8,468	242,165	5.17
⌀(2–12)			8,437	221,721	5.20

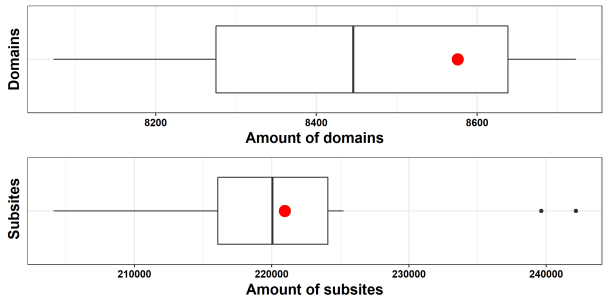


Figure 4: Number of domains and subsites visited in our measurements (M#1–M#12). The dots represent M#1.

Table 3: Overview of the measured graph structures (with and without isolated nodes) in terms of observed nodes (companies) and connections between them. The relative percentages refer to M#1.

ID	Number of nodes total		without iso.		connections total (without iso.)	
M#1	12,304	—	566	—	842	—
M#2	10,380	-15.6 %	381	-32.7 %	499	-40.7 %
M#3	9,811	-20.3 %	355	-37.3 %	447	-47.9 %
M#4	10,265	-16.6 %	347	-38.7 %	422	-49.9 %
M#5	9,997	-18.8 %	316	-44.2 %	362	-57.0 %
M#6	8,348	-32.2 %	293	-48.2 %	339	-59.7 %
M#7	10,365	-15.8 %	361	-36.2 %	426	-49.4 %
M#8	10,192	-17.2 %	355	-37.3 %	416	-50.6 %
M#9	10,466	-14.9 %	395	-30.2 %	430	-48.9 %
M#10	10,601	-13.8 %	302	-46.6 %	316	-62.5 %
M#11	9,647	-21.6 %	329	-63.4 %	373	-55.7 %
M#12	11,240	-8.7 %	348	-38.5 %	419	-50.2 %
⌀(2–12)	10,119	-17.8 %	344	-41.2 %	404	-52.0 %

5.1 Third-Party Sharing Ecosystem

The data of each measurement was processed and sorted to construct a graph that represents embedded third parties and information sharing networks (see Section 4 and Table 3). All graphs are undirected. Figure 5 visualizes graph plots of the first two measurements. Nodes represent companies and edges represent ID syncing between the companies. Therefore, the nodes reflect the total number of third parties embedded in websites and could potentially collect and share personal data. A decrease in the number of nodes means that first parties embed—directly or indirectly—less third parties (e. g., less trackers or companies participate in the ad bidding process). The amount of edges reflects the number of companies syncing IDs. A smaller number of edges means that fewer companies participate in the sharing economy. The most dominant important node is representing *Google*. Other important nodes represent companies such as *AppNexus*, *Amazon*, or *Oracle*.

Figures 6a and 6b show the number of nodes and edges per measurement. The y-axis represents the number of nodes or connections and the x-axis represents the calendar weeks (CW). The thick light gray dot on the left is the first measurement M#1, in CW 20, *before* the GDPR came into effect, and the dark gray dots represent the other measurements (M#2 to M#12). We performed two types of linear regression analysis including the measurement, one before the GDPR took effect y_{pre}(gray dotted line) and one excluding it, y_{post} (black dashed line).

We chose a linear regression because a nonlinear regression for the number of measuring points and values could lead to overfitting. Moreover, the Pearson (nodes pre: 0.3, nodes post: -0.0; sync pre: -0.5, sync post: -0.6) and Spearman (nodes pre: 0.3, nodes post: 0.0; sync pre: -0.6, sync post: -0.7) coefficients are close to each other, indicating that linear regression is appropriate for our purpose. Comparing both trends, we see a significant difference in the slope of the regression lines.

To confirm that the number of embedded third parties over all websites between M#2–M#12 is statistically significantly different from M#1, we calculate the confidence interval (99 % confidence) for the prediction of the previous curve for the pre-GDPR measurement on the basis of the values without the value of measurement M#1. If the value of our pre-GDPR measurement is outside the confidence interval, we confirm that by the time of the introduction of the GDPR, the number of nodes has decreased.

The result is 7,151 as the lower confidence limit and 11,774 as the upper confidence limit (see the red interval in Figure 6a). With a value of 12,304, the first measurement is barely outside the interval. Thus, we see evidence that the amount of parties used in M#1 is independent of the number of parties observed in the remaining EU measurements. We need to be careful in the interpretation of these numbers as it is a matter of an effect of the GDPR and not directly about the GDPR itself. The strength of the effect is rather small, since the value of M#1 lies only barely outside the interval.

As shown in Table 2, the amount of third parties per website stays more or less stable across all measurements, while Figure 6a shows a drop of third parties used from M#1 to M#2. However, Table 2 lists domain averages and Figure 6a shows companies aggregated over all domains. The overall decrease is in line with previous work that found that websites tended to switch to larger ad networks (e. g., *Google* or *Facebook*) when the GDPR took effect [8]. Thus, it is reasonable that the absolute number of observed companies drops (smaller companies disappear), while the total amount of third parties stays stable. We discuss the measured effects on companies active in the ecosystem in Section 6.

Before the GDPR enforcement, the graph M#1 contained 12,304 nodes, 11,738 of which are isolated. Isolated nodes have no connection to another node and represent third-party companies that are embedded into websites but do not perform cookie syncing (e. g., a JavaScript library). Overall, the number of third parties, isolated or not, decreases over the course of our study. However, without the pre-GDPR measurement the trend of embedded thrid parties is slowly rising. All further findings *exclude* the isolated nodes (i. e., we only analyzed the nodes that engage in cookie syncing).

Figure 6b shows the number of ID sharing connections. Of particular interest is the reduction of syncing relations by about 40 % over the course of our measurement—in terms of the number of direct syncing connections. The corresponding linear regression analysis confirms that both trends with (y_{pre}, gray line) and without (y_{post}, black line) the pre-GDPR measurement are both decreasing to different extents.

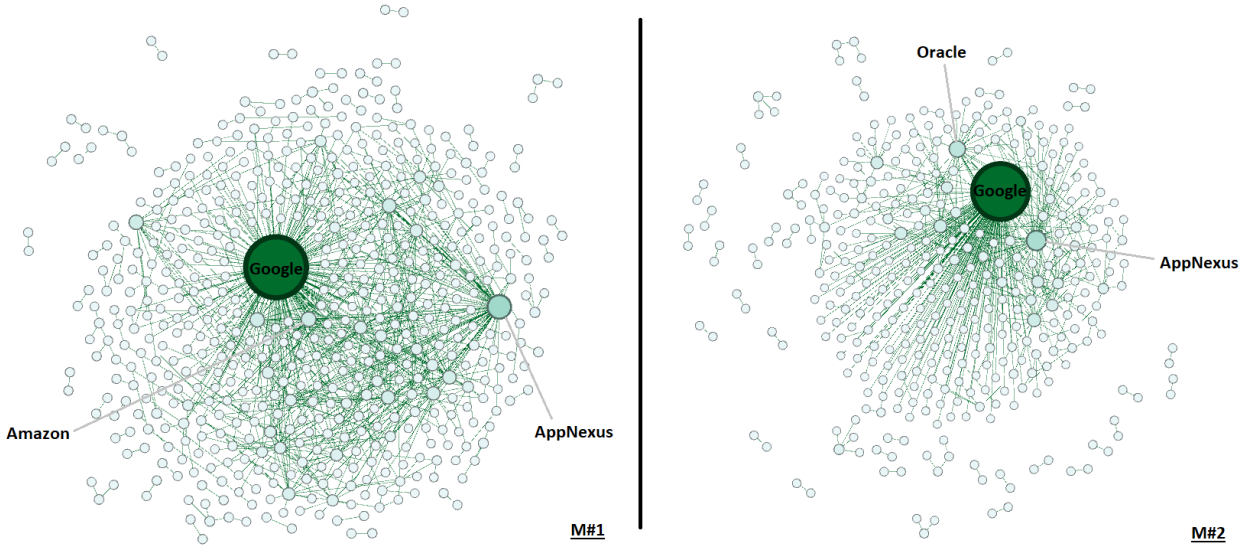
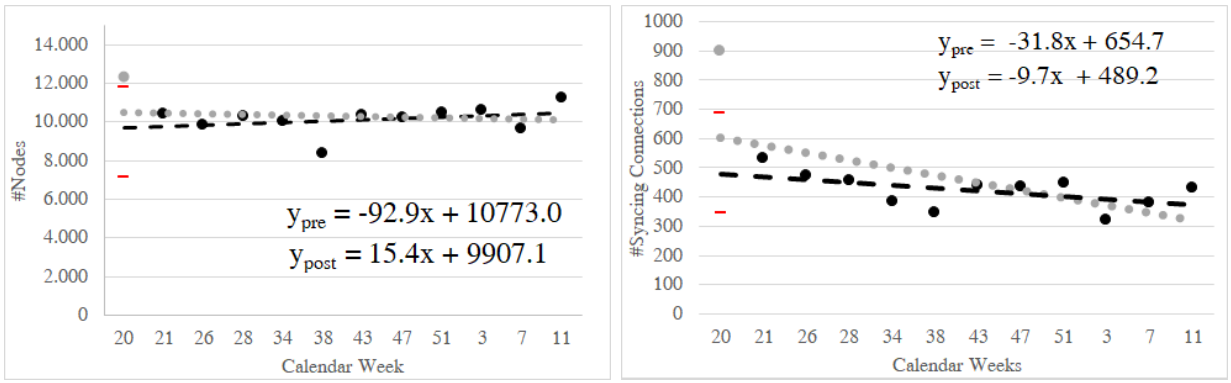


Figure 5: The graphs demonstrate the change of syncing connection between our pre-GDPR measurement on May, 19 2018 (M#1, left) and the measurement right after the GDPR went into effect on May 25, 2018 (M#2, right). A reduction of nodes and edges is visible. The weight, calculated by the PageRank algorithm, of the individual nodes in the graph is represented by the strength of the color and size of the node (the darker and bigger, the more important). The importance of the edges is also quantified by the color (the darker, the more important). Additionally, the three most significant nodes are labeled.



(a) Number of third parties per measurement

(b) Number of syncing connections per measurement

Figure 6: Regression lines of our measurements including the pre-GDPR measurement (gray) and excluding it (black). The red dashes represent the confidence interval (99 % confidence) of the prediction for the pre-GDPR measurement point based on all post-GDPR measurements.

To test if there is a statistically significant difference in ID syncing activities between M#1 and the remaining EU measurements, we again calculate a 99 %-confidence interval for the prediction of the curve for the pre-GDPR measurement on the basis of the values without the pre-GDPR measurement. The pre-GDPR measurement value (898) is outside the interval (347 lower limit, 686 upper limit), thus we see strong evidence that in M#1 and the remaining measurements different levels of ID syncing occurred. In this case, the strength of the effect is more obvious than with the nodes before.

Furthermore, we compared the linear regression lines including (y_{pre}, dotted gray lines) and excluding (y_{post}, dashed black lines) the pre-GDPR measurement. In both cases, the slopes are lower which indicates that the drop between the first and second measurement is significantly larger than in the following weeks but is part of a general trend towards fewer third parties that also sync less.

Table 4 provides an overview of the connections within the graphs, excluding the isolated nodes. To measure whether the effects on the number of third parties and syncing are independent,

we separated the graphs into connected components. Each connected component represents a subgraph in which nodes are connected to each other by paths. M#1 has 59 components, with the largest component containing 429 nodes. The percent values reflect the reduction and always refer to the initial graph M#1, so the number of components is reduced from M#1 to M#12 by a maximum of around 56 % (M#6). Another difference is the size of the largest component, which is reduced by up to 55 % (M#10). However, the median component size remains stable at around two throughout all measurements. This indicates that overall components were not affected by the disappearing connections. However, the number of components did drop.

Similarly, the algebraic connection is a measure for the number of nodes and the number of connections between the nodes within the graph. This value can be interpreted as the robustness of the graph with regard to the connections. The lower the value, the fewer connections are present. The values of the algebraic connection vary between positive 25 % and negative 60 % compared to the initial measurement. The evaluation shows that the total number of links in the graph fluctuates, but numbers are similar comparing the first and the last measurement (-0.51 %). Although individual measurements vary due to the internal structure of the ecosystem over the course of our measurements, we did not measure a significant effect on the structure of our graphs over time.

The reduction in the number of edges and nodes both follow an overall downward trend: Fewer third parties are present in the ecosystem and these share fewer IDs (see Figures 6a and 6b). However, over the month following the introduction of the GDPR, the number of nodes slightly increases again, whereas the number of edges continues to decrease. Therefore, the number of nodes can theoretically be represented by a quadratic function.

Table 4: Overview of connected components (CP) in the measured graphs (M#1–M#12) and the shift after the GDPR took effect.

ID	Components		Connectivity largest CP		algebraic conn.	
M#1	59	—	429	—	0.1187	—
M#2	38	-35.6 %	296	-31.0 %	0.1494	+25.9 %
M#3	37	-37.3 %	269	-37.3 %	0.1071	-9.8 %
M#4	30	-49.2 %	277	-35.4 %	0.0994	-16.3 %
M#5	37	-37.3 %	235	-45.2 %	0.0818	-31.1 %
M#6	26	-55.9 %	225	-47.6 %	0.0469	-60.5 %
M#7	38	-35.6 %	268	-37.5 %	0.1146	-3.5 %
M#8	38	-35.6 %	275	-35.9 %	0.0488	-58.9 %
M#9	47	-20.3 %	284	-33.8 %	0.0479	-59.6 %
M#10	45	-23.7 %	193	-55.0 %	0.1181	-0.5 %
M#11	36	-34.0 %	247	-42.4 %	0.0654	-44.9 %
M#12	35	-40.7 %	267	-37.8 %	0.0829	-44.5 %
⌀(2–12)	37	-37.4 %	258	-39.9 %	0.0875	-27.2 %

Comparing the results from our crawls conducted in Europe with our two reference measurements from US-based IP addresses, we observed that the amount of cookie syncing for website visits from the USA is about 15 % higher than the amount measured at a similar points in time from the EU (CW43 and CW5—which where conduct one week prior to the US measurements). Furthermore, we found that there are more connected nodes in the US measurements (+33 %) and that there are less components (-22 %) but the existing components are larger (+9 %) and more connected (+58 %). However, we observed less nodes in total (-12 %). Hence, in our US measurements we observed less third parties in general but these sync private data more extensively and are more connected with each other.

Table 5 presents the general graph characteristics of our conducted measurements (M#1-M#12). The longest possible distance between two nodes (i. e., the diameter), modularity and medium degree of the graphs remains more or less stable. Nevertheless, the number of communities is reduced from 69 in M#1 to 50 communities in M#2 and 47 communities in M#3, and even 34 communities in M#6. Note that the values of communities and the values of modularity may vary due to the algorithm used to determine the values. We use the software `Gephi 0.9.2` [5] to compute the communities and modularity. The average clustering coefficient shows a decrease. The average distance between node pairs in the graph indicates the average path length. These values do not change much across the course of all our measurements. This indicates that the underlying ecosystem remains unchanged.

Table 5: Characteristics of our graphs without isolated nodes.

ID	diameter	median degree	modularity	⌀ clustering coeff.	⌀ path length	comm.
M#1	9	2.98	0.58	0.23	3.13	69
M#2	8	2.61	0.61	0.18	3.10	50
M#3	8	2.52	0.64	0.18	3.23	47
M#4	9	2.43	0.66	0.15	3.35	42
M#5	10	2.29	0.65	0.16	3.19	47
M#6	10	2.31	0.72	0.07	3.93	34
M#7	9	2.36	0.72	0.07	3.50	45
M#8	11	2.34	0.67	0.08	3.58	50
M#9	12	2.18	0.71	0.07	3.73	58
M#10	8	2.09	0.72	0.04	3.46	55
M#11	9	2.27	0.70	0.05	3.68	36
M#12	10	2.41	0.67	0.05	3.66	35
⌀(2–12)	9	2.35	0.68	0.10	3,49	46

5.2 Connections of Third Parties

To get a better understanding of the described effects on the tracking ecosystem, we analyze the structure of the measured third-party graphs. We look at the degree of each node and classify them based on the number of *direct* and *indirect* partners. Primary partners are those where a direct syncing relation was observed while secondary partners are those with a higher degree of separation. We classified third parties (nodes) into three categories: (1) nodes with predominately direct (primary) partners, (2) nodes with only one partner but a large number of secondary partners, and (3) nodes with a rather balanced amount of primary and secondary partners. We labeled a node "central" if it has four times more primary partners than secondary partners, "outer" if it has four times more secondary

partners than primary partners, and "balanced" otherwise. Our data set contains 21 central nodes and 30 balanced nodes. The remaining nodes in the graph are end nodes in a star.

The majority networks of cooperating third parties are arranged in star topologies. They have one central point with many primary syncing connections to partners (e. g., *Google*), but these partners rarely sync with additional partners. Other nodes with many secondary partners have few primary partners (often just 1), who are the central point of a star. Thus, these companies are connected to all outer nodes of the star as secondary partners. Nodes with a balanced amount of primary and secondary partners do not have any other special characteristics.

We also analyzed the effects of the mean betweenness centrality between the pre-GDPR measurement and the post-GDPR measurements. The betweenness centrality is an index to measure how many shortest paths in a graph include a node. The higher the betweenness centrality of a node, the higher the amount of information that flows through this node. For example, a central node in a star topology would have a high betweenness centrality index, because it is the center of the star, while the outer nodes would have a betweenness centrality index of zero (they are only the start/end of the shortest path but never have multiple edges). In contrast to the degree of a node, the betweenness centrality can been seen as factor measuring the links between two star typologies. Hence, a high betweenness centrality score shows that a node connects different syncing communities (i. e., serves as a "bridge"). We computed the betweenness centrality index using the `NetworkX` Python package [23].

Similar to our syncing connection regression, we performed a linear regression of the mean betweenness centrality and found a statistically significant ($\alpha = 0.01$ with p-value $< .001$) decrease in the betweenness centrality. In extreme cases, the betweenness centrality dropped by up to 60 % (mean 30 % SD: 11 %). An overview of the betweenness centrality properties of our measured graphs is given in Table 6 where all graphs have a median and minimum betweenness centrality of zero. We used the 75 % quantile of the betweenness centrality of all nodes observed in M#1, 5.87, as a reference value to illustrate the change of betweenness centrality over time.

In line with the findings that the amount of syncing connections decreases, the mean/max betweenness centrality also decreases. Furthermore, the amount of well-connected nodes (b/c $\geq$ 5.87 in our case) and connected nodes decreases which also means that fewer nodes sync IDs with each other.

The result of fewer companies participating in the ID sharing has different effects on the importance of different nodes—in terms of sharing connections and the information flowing through the nodes. The betweenness centrality of the most important node, *Google*, decreases by around 36 % while other nodes actually gain (e. g., *Oracle* (71 %) or *MediaMath* (24 %)) in betweenness centrality. However, in absolute numbers *Google* is still the dominant node in our graph. Overall, 43 companies gained betweenness centrality, 78 lost betweenness centrality ($\leq$ 50 %), and the betweenness centrality of 31 companies was decreased significantly by more than 50 %. These numbers only include companies that were observed in M#1 and at least two other EU measurements. The nodes gaining betweenness centrality are mostly small companies with initially low betweenness centrality scores of less than 5.87 (37).

Regarding the classification of a node, we found that, due to the star-like topologies, that "central" nodes have high betweenness centrality scores and "outer" nodes have low (or zero) betweenness centrality scores. In our scenario, the betweenness centrality can be seen as a metric how prevalent a company is in the syncing ecosystem. Thus, these companies are connected to all outer corners of the star as secondary partners. However, we did not see a paradigm change how companies sync user IDs. Overall, the degree distribution in our measured graphs did not vary a lot between all graphs (see Appendix A), but the total amount of links dropped by 23 %.

These observations are in line with the results of our previous observations that the general structure (or business practices) within the ecosystem did not change after the GDPR became effective, but we have shown that ID syncing dropped significantly. Over the course of our study, we observed that the number of primary partners of most companies continuously decreased by up to 40 % (83 less primary partners). Five companies became isolated and only two companies gained primary partners. With respect to secondary connections, we see a fluctuation of partners. This can be explained by the fact that adding one primary partner, who might be the center of another star, can lead to a significant number of additional secondary partners (sometimes hundreds of secondary partners).

Our results also show that embedding one third party into a website puts users at risk that their data gets shared with hundreds of companies. This leads to the problem that users cannot verify who received a copy of their data, which leads to the question how service providers can ensure that data is deleted upon request. Previous work conducted prior to the GDPR has found that an ID is synced with 3.5 partners on average [38]. Our measurements have shown that the average amount of ID syncing partners might not be a good metric to assess ID syncing due to the star-like topology, rather an in depth graph analysis is necessary. Aside from the one dominating star, with *Google* as a central point, we observe many smaller networks that share IDs with each other. This is in line with our observation of the communities in the graph (see Table 4) and public announcements of companies to build tracking infrastructures besides *Google* or *Facebook* [35].

5.3 Case Studies

Only 70 companies were observed in all 11 measurements. Most of those are prominent companies that offer multiple services (e. g., *Google* or *Oracle*). A summary of these companies and how they evolved over time is given in Appendix B. We found 20 companies (approx. 3 %) that had shared data before May 28, 2018 and were not observed in any of the consecutive EU measurements, but still appeared in our US measurements. Manual inspection of these services showed that some had announced they were discontinuing business in the European Union or changed their business model. For example, one website stated: "*Currently, XX does not provide any services in the European Economic Area (EEA), service will be resumed once we feel that we are able to comply with the GDPR criteria.*". Two other companies notified their customers that they were required

Table 6: Betweenness centrality properties of graphs and the changes of the most central nodes over time.

ID	mean		sd		max		b/c = 0		b/c < 5.87		b/c ≥ 5.87	
M#1	345	—	3,022	—	68,852	—	395	—	29	—	142	—
M#2	241	-30 %	1,821	-40 %	33,978	-51 %	262	-34 %	15	-48 %	104	-27 %
M#3	227	-34 %	1,507	-50 %	26,277	-62 %	251	-36 %	15	-48 %	88	-38 %
M#4	259	-25 %	1,711	-43 %	29,197	-58 %	235	-41 %	12	-59 %	100	-30 %
M#5	191	-45 %	1,336	-56 %	22,043	-68 %	228	-42 %	13	-55 %	75	-29 %
M#6	253	-27 %	1,153	-62 %	16,496	-76 %	186	-53 %	14	-52 %	93	-47 %
M#7	248	-28 %	1,524	-50 %	26,523	-61 %	244	-38 %	22	-24 %	95	-33 %
M#8	274	-21 %	1,678	-44 %	28,886	-58 %	254	-36 %	8	-72 %	93	-35 %
M#9	278	-19 %	1,715	-43 %	32,135	-53 %	285	-28 %	18	-38 %	92	-35 %
M#10	151	-56 %	862	-71 %	13,525	-80 %	214	-46 %	21	-27 %	67	-53 %
M#11	248	-28 %	1,375	-55 %	21,714	-68 %	234	-41 %	17	-41 %	78	-45 %
M#12	271	-22 %	1,562	-48 %	25,832	-62 %	246	-38 %	14	-52 %	88	-38 %
⌀(2 – 12)	240	-30 %	1,477	-51 %	25,146	-63 %	240	-40 %	15	-48 %	88	-38 %

to adopt a technology based on consent management platforms[2] (CMP): "*But please keep in mind, if you do not comply with GDPR, then XXX (and many other ad tech partners) will not be able to monetize any of your EU traffic.*" Since our data collection setup did not automatically give consent, these companies are likely compliant with the new standards and stopped sharing data without consent. Aother company announced in early 2018 that it was refocusing its business towards contextual advertising, where ads are based on the content of a website and not the profile of the user visiting the website. However, for the majority of companies, we did not find GDPR-related information, but it is possible that they quietly retreated from the European market, without publicly explaining that their services cannot be made compliant.

Overall, our data shows that companies share data with a smaller number of partners that they did in early 2018 which is in line with other studies that have shown that the reach of smaller companies has decreased, while tracking by the market leaders has increased [22]. An alternative explanation for our results is that companies changed how they exchange IDs. Our measurement approach (see Section 4) relies on ID syncing that can be observed on the client side. Therefore, it is possible that a shift towards server-side ID syncing is taking place that cannot be studied with current methods. Previous work found that *Google* is one of the beneficiaries of the GDPR, as the number of websites that embed one of their services [22] increased. Regarding the amount of information flowing through nodes, in terms of ID sharing, we cannot confirm these findings. According to our measurements, *Google* and others, lost importance in that regard while other nodes, especially *Oracle*, gained importance. However, in total *Google* is still the leading company. Our results are not contradicting findings of previous work but are complementary: Previous work has shown that *Google* has increased its reach (numbers of websites directly embedding their services) and our results show information flowing through *Google* (by other third parties) is reduced.

[2] See https://advertisingconsent.eu/ for details.

6 CONCLUSION AND LIMITATIONS

Our measured third-party graphs represent only a small subset of the real third-party relations of a website. A website might detect that a crawler is visiting it and embed different objects or none at all, even though we tried to mask our crawler. Aside from scrolling and mouse jiggling, we do not interact with the websites, which might also influence our results because some third parties might only be embedded if a user performs a specific task (e. g., if the user starts a purchase process, a third party might be embedded to handle the credit card payment). We did not interact with any cookie consent banners present on the visited websites. Therefore, we might not have observed all cookie syncing attempts and our results can be seen as a lower bound.

However, previous works found that cookie banners often do not work as expected [41], do not offer opt-out choices while instead assume opt-in [55], and showed that the used consent libraries do not meet other GDPR requirements [15]. In addition, Utz et al. [55] have also shown that the majority of users does not interact with cookie consent notices, similar to our approach.

After our first measurement, conducted before the GDPR took effect, we observed a statistically significant drop in ID sharing connections within the online advertising ecosystem. It is likely that the change is related to the GDPR, which imposed stricter rules on data sharing and allows data protection authorities to fine non-compliant companies. However, we cannot exclude other factors that could have caused a change in the ecosystem, like the adoption of a new technology for ID sharing. We do not measure syncing attempts other than those using requests (e. g., based on IP addresses or device fingerprints [56]). To the best of our knowledge, cookie syncing is still the most prevalent way to share user identifiers. It is also possible that the change had nothing to do with the GDPR and are purely coincidental, but the public debates on the topic around that time suggest a relation. Note that we do not attempt to measure when companies share all of its collected personal data with another company at once (e. g., *Facebook* sharing all of their collected data with other companies [49]) but rather want to explore data sharing happening in real time on the browser level.

While the number of companies and the number of direct connections decreased around May 2018, the trend stabilized and the number of third parties has increased since then. This could be an indicator that some websites only temporally stopped the use of some services but over time took the necessary steps to use these services again under GDPR (e. g., signing data processing agreements). This observation is in line with other studies [19, 22]. Regarding the structure of the measured graphs, we did not see a significant change in the ecosystem. This hints that companies did not change their business practices but are more cautious when it comes to the processing of personal data. The GDPR might have caused a disruption in the online advertising ecosystem as ID syncing—an important part of the ecosystem—significantly decreased, but neither revolutionized it as the structures remained intact nor did it dispatch the ecosystem as some industry-related groups had pessimistically forecasted [34, 46]. More importantly, the effects on Internet users' privacy might be negative as fewer companies continue to be present on more websites, increasing their possibilities to create profiles. The results indicate that the characteristics of the ad ecosystem did not change during the course of our study. Cookie syncing is still used in practice, but its extent is significantly reduced and still declining.

In contrast to previous work (see Section 3), we found statistically significant changes in the online advertising ecosystem around the GDPR enforcement date. Other work focused on embedded third parties [43] or more specifically tracking companies [15] but could not measure a direct impact. However, this does not contradict other results as we also found that the ecosystem, in general, did not change. To a greater degree, our work shows that the effects of the GDPR might not be directly measurable in all aspects of the online ecosystem but in-depth analysis is needed to get a better understanding of the effects of such complex legislation in a complex environment. Future work should investigate how different companies actually implemented their data sharing practices either by actively making use of the *right to access*, granted by the GDPR, or by conducting expert interviews to understand the reasons why companies changed their practices in some areas (e. g., data sharing) but apparently not in others (e. g., tracking). Recent fines [47] and ongoing legal complaints [6] for lack of transparency indicate that such aspects need to be studied in more detail.

ACKNOWLEDGMENTS

This work was partially supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia (MKW grants 005-1703-0021 "MEwM" and Research Training Group NERD.nrw). We would like to thank the anonymous reviewers for their valuable feedback and Christine Utz at Ruhr University Bochum for her efforts proofreading this work. Any findings, conclusions, opinions, or recommendations stated in this work are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. 2014. The Web Never Forgets—Persistent Tracking Mechanisms in the Wild. In *Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS'14)*. ACM Press, New York, New York, USA, 674–689.

[2] Alexa.com. 2018. Top Sites for Countries. https://www.alexa.com/topsites/countries Accessed: 2019-02-05.

[3] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. 2016. Tracing Information Flows Between Ad Exchanges Using Retargeted Ads. In *Proceedings of the 25th USENIX Security Symposium (USENIX Sec'16)*. USENIX Association, Austin, TX, 481–496.

[4] Muhammad Ahmad Bashir and Christo Wilson. 2018. Diffusion of User Tracking Data in the Online Advertising Ecosystem. *Proceedings on Privacy Enhancing Technologies* 4 (2018), 85–103.

[5] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*. AAAI, San Jose, United States, 361–362.

[6] Brave Browser. 2019. Update on GDPR complaint (RTB ad auctions). https://www.brave.com/blog/update-rtb-ad-auction-gdpr/ Accessed: 2019-05-09.

[7] Randolph E. Bucklin and Catarina Sismeiro. 2003. A model of web site browsing behavior estimated on clickstream data. *Journal of Marketing Research* 40, 3 (2003), 249–267.

[8] Cliqz. 2018. GDPR - What happened? https://whotracks.me/blog/gdpr-what-happened.html Accessed: 2019-10-05.

[9] Cliqz. 2018. WhoTracks.me Data – Tracker database. https://whotracks.me/blog/gdpr-what-happened.html Accessed: 2019-04-24.

[10] Commission Nationale de l'Informatique et des Libertés. 2019. Deliberation of the Restricted Committee SAN-2019-001 of 21 January 2019 pronouncing a financial sanction against GOOGLE LLC. https://www.cnil.fr/sites/default/files/atoms/files/san-2019-001.pdf Accessed: 2019-10-05.

[11] CONSENT project. 2017. CONSENT Report Summary. https://cordis.europa.eu/result/rcn/140471_en.html Accessed: 2019-02-05.

[12] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. 2019. Measuring Cookies and Web Privacy in a Post-GDPR World. In *Proceedings of the 2019 Conference on Passive and Active Measurement (PAM'19)*. Springer-Verlag, Cham, 258–270.

[13] Data Protection Working Party. 2010. Opinion 2/2010 on online behavioural advertising. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp171_en.pdf Accessed: 2019-10-05.

[14] Martin Degeling and Jan Nierhoff. 2018. Tracking and Tricking a Profiler: Automated Measuring and Influencing of Bluekai's Interest Profiling. In *Proceedings of the 2018 ACM Workshop on Privacy in the Electronic Society (WPES'18)*. ACM Press, New York, New York, USA, 1–13.

[15] Martin Degeling, Christine Utz, Christoper Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proceedings of the 2019 Symposium on Network and Distributed System Security (NDSS'19)*. Internet Society, San Diego, California, USA.

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

[17] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC

[18] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. 2016. Tracking Personal Identifiers Across the Web. In *Proceedings of the 2016 Conference on Passive and Active Measurement (PAM'16)*, Thomas Karagiannis and Xenofontas Dimitropoulos (Eds.). Springer, Cham, 30–41.

[19] FutureScot. 2018. An unexpected benefit of GDPR; it makes the web much faster. http://futurescot.com/an-unexpected-benefit-of-gdpr-it-makes-the-web-much-faster/ Accessed: 2019-02-05.

[20] Global Stats. 2019. Screen Resolution Stats. http://gs.statcounter.com/screen-resolution-stats Accessed: 2019-02-05.

[21] Roberto Gonzalez, Lili Jiang, Mohamed Ahmed, Miriam Marciel, Ruben Cuevas, Hassan Metwalley, and Saverio Niccolini. 2017. The cookie recipe: Untangling the use of cookies in the wild. In *Proceedings of the 2017 Network Traffic Measurement and Analysis Conference (TMA'17)*. IEEE, Piscataway, NJ, 1–9.

[22] Björn Greif. 2018. Study: Google is the biggest beneficiary of the GDPR. https://cliqz.com/en/magazine/study-google-is-the-biggest-beneficiary-of-the-gdpr Accessed: 2019-02-05.

[23] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, Gaël Varoquaux, Travis Vaught, and Jarrod Millman (Eds.). SciPy, Pasadena, CA USA, 11 – 15.

[24] Paul De Hert, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, and Ignacio Sanchez. 2018. The right to data portability in the GDPR: Towards user-centric interoperability of digital services. *Computer Law & Security Review* 34, 2 (2018), 193–203.

[25] IAB Europe. 2017. European Digital Advertising market has doubled in size in 5 years. https://www.iabeurope.eu/research-thought-leadership/resources/iab-europe-report-adex-benchmark-2017-report/ Accessed: 2019-10-05.

[26] Interactive Advertising Bureau. 2017. Internet Advertising Revenue Report. https://www.iab.com/wp-content/uploads/2018/05/IAB-2017-Full-Year-Internet-Advertising-Revenue-Report.REV2_.pdf Accessed: 2019-10-05.
[27] Internet World Stats. 2018. Top 20 countries with the highest number of internet users. https://www.internetworldstats.com/top20.htm Accessed: 2019-02-05.
[28] Luca Invernizzi, Kurt Thomas, Alexandros Kapravelos, Oxana Comanescu, Jean-Michel Picod, and Elie Bursztein. 2016. Cloak of Visibility: Detecting When Machines Browse a Different Web. In *Proceedings of the 2016 IEEE Symposium on Security and Privacy (S&P'16)*. IEEE, Piscataway, NJ, 743–758.
[29] IPLocation. 2019. Where is Geolocation of an IP Address? https://www.iplocation.net/ Accessed: 2019-02-05.
[30] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M. Pujol. 2018. WhoTracks.Me: Monitoring the online tracking landscape at scale. *CoRR* abs/1804.08959 (2018). arXiv:1804.08959 http://arxiv.org/abs/1804.08959
[31] Mohammad Taha Khan, Joe DeBlasio, Geoffrey M. Voelker, Alex C. Snoeren, Chris Kanich, and Narseo Vallina-Rodriguez. 2018. An Empirical Analysis of the Commercial VPN Ecosystem. In *Proceedings of the 2018 Internet Measurement Conference (IMC'18)*. ACM Press, New York, USA, 443–456.
[32] Timothy Libert. 2018. An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conference Committee, Republic and Canton of Geneva, Switzerland, 207–216.
[33] Timothy Libert, Lucas Graves, and Rasmus Kleis Nielsen. 2018. *Changes in third-party content on European News Websites after GDPR.* Technical Report. Reuters Institute for the Study of Journalism, Oxford, UK. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-08/Changes%20in%20Third-Party%20Content%20on%20European%20News%20Website%20after%20GDPR_0_0.pdf Accessed: 2019-10-05.
[34] MyCustomer. 2018. Will GDPR kill the third-party data market? https://www.mycustomer.com/marketing/data/will-gdpr-kill-the-third-party-data-market Accessed: 2019-05-09.
[35] Gabriel Nunes. 2018. Adobe is helping some 60 companies track people across devices. https://www.neowin.net/news/adobe-is-helping-some-60-companies-track-people-across-devices Accessed: 2019-10-05.
[36] Lukasz Olejnik and Claude Castelluccia. 2014. *To bid or not to bid? Measuring the value of privacy in RTB.* Technical Report. INRIA, Grenoble. 23 pages.
[37] PageFair. 2017. The 3 biggest challenges in GDPR for online media & advertising. https://pagefair.com/blog/2017/gdpr-3-deep-challenges/ Accessed: 2019-10-05.
[38] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. 2018. Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. *CoRR* abs/1805.10505 (2018).
[39] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. 2018. The Cost of Digital Advertisement: Comparing User and Advertiser Views. In *Proceedings of the 2018 World Wide Web Conference (WWW'18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1479–1489.
[40] John W Ratcliff and David E Metzener. 1988. Pattern matching: The Gestalt Approach. *Dr Dobbs Journal* 13, 7 (1988), 46.
[41] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. 2019. Can I Opt Out Yet?: GDPR and the Global Illusion of Cookie Control. In *Proceedings of the 2019 ACM Symposium on Information, Computer and Communications Security (AsiaCCS'19)*. ACM Press, New York, New York, USA, 340–351.
[42] Klaus Schwab, Alan Marcus, Justin Rico Oyola, William Hoffman, and Michele Luzi. 2011. *Personal data: The emergence of a new asset class.* Technical Report. World Economic Forum.
[43] Jannick Kirk Sørensen and Sokol Kosta. 2019. Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. In *Proceedings of the 2019 World Wide Web Conference (WWW'19)*. International World Wide Web Conferences Committee, Republic and Canton of Geneva, Switzerland, 1590–1600.
[44] Oleksii Starov and Nick Nikiforakis. 2017. Extended Tracking Powers. In *Proceedings of the 26th World Wide Web Conference (WWW '17)*. ACM Press, New York, New York, USA, 1481–1490.
[45] Tech Blog. 2019. Most Common User Agents. https://techblog.willshouse.com/2012/01/03/most-common-user-agents/ Accessed: 2019-02-05.
[46] The Drum. 2017. The day after tomorrow: when adblockers and GDPR kill all adtech and martech. https://www.thedrum.com/opinion/2017/10/17/the-day-after-tomorrow-when-ad-blockers-and-gdpr-kill-all-adtech-and-martech Accessed: 2019-05-09.
[47] The Guardian. 2019. Google fined record £44m by French data protection watchdog. https://www.theguardian.com/technology/2019/jan/21/google-fined-record-44m-by-french-data-protection-watchdog Accessed: 2019-05-09.
[48] The New York Times. 2014. The Slow Death of 'Do Not Track'. http://www.nytimes.com/2014/12/27/opinion/the-slow-death-of-do-not-track.html Accessed: 2019-10-05.
[49] The New York Times. 2018. 5 Ways Facebook Shared Your Data. https://www.nytimes.com/2018/12/19/technology/facebook-data-sharing.html Accessed: 2019-02-05.
[50] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 2019. 4 Years of EU Cookie Law: Results and Lessons Learned. *Proceedings on Privacy Enhancing Technologies* 2019, 2 (2019), 126–145. https://content.sciendo.com/view/journals/popets/2019/2/article-p126.xml
[51] TRUSTe and Harris Interactive. 2011. Consumer Research Results - Privacy and Online Behavioral Advertising. https://www.eff.org/files/truste-2011-consumer-behavioral-advertising-survey-results.pdf Accessed: 2019-02-05.
[52] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2019. "Your Hashed IP Address: Ubuntu."–Perspectives on Transparency Tools for Online Advertising. In *Proceedings of the 2019 Anual Computer Security Applications Conference (ACSAC'19)*. ACM Press, New York, New York, USA.
[53] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2019. A Study on Subject Data Access in Online Advertising After the GDPR. In *Proceedings of the 14th Workshop on Data Privacy Management (DPM'19)*. Springer-Verlag, Cham, 61–79.
[54] Tobias Urban, Dennis Tatang, Thorsten Holz, and Norbert Pohlmann. 2018. Towards Understanding Privacy Implications of Adware and Potentially Unwanted Programs. In *Proceedings of the 2018 European Symposium on Research in Computer Security (ESORICS'18)*. Springer-Verlag, Cham, 449–469.
[55] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM Conference on Computer and Communications Security (CCS'19)*. ACM Press, New York, USA, 973–990.
[56] Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy. 2018. FP-STALKER: Tracking Browser Fingerprint Evolutions. In *Proceedings of the 39th IEEE Symposium on Security and Privacy (S&P'18)*. IEEE, San Francisco, United States, 728–741.
[57] Zachary Weinberg, Shinyoung Cho, Nicolas Christin, Vyas Sekar, and Phillipa Gill. 2018. How to Catch when Proxies Lie: Verifying the Physical Locations of Network Proxies with Active Geolocation. In *Proceedings of the 2018 Internet Measurement Conference (IMC'18)*. ACM Press, New York, USA, 203–217.
[58] What Is My IP Address. 2019. Where is Geolocation of an IP Address? https://whatismyipaddress.com/ Accessed: 2019-02-05.
[59] Yong Yuan, Feiyue Wang, Juanjuan Li, and Rui Qin. 2014. A survey on real time bidding advertising. In *Proceedings of the 2014 Conference on Service Operations and Logistics, and Informatics (SOLI'14)*. IEEE, Piscataway, NJ, 418–423.

A GRAPH CHARACTERISTICS

Figure 7 shows the density of degrees of all nodes in our third-party graphs (normalized). We did not measure a significant shift in the distribution of links, however, the total number of links shrunk around 23 %.

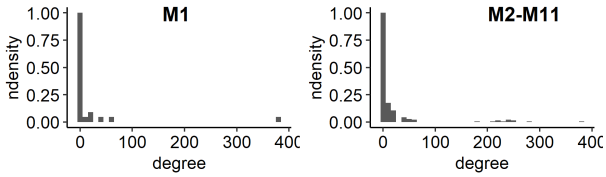


Figure 7: Overview of the distribution of the measured degrees of all nodes (excluding isolated notes).

B ANALYSIS CORPUS CLASSIFICATION

Table 7 lists direct syncing partners of the companies most often observed in our experiments. Furthermore, the table shows the node types (i. e., outer, balanced, center, and isolated). The rows 'Remaining Nodes' show the mean syncing relations from all third parties that are not present in our analysis corpus. To increase readability, we only focus on the first four measurements.

For most third parties, the number of direct partners is reduced over the course of our measurements. The biggest reduction is

attributed to *Google* and the only exception is *Adform*. Interestingly, the number of direct connections from M#2 to M#3 increases again in some cases, for example, for *Criteo*. In principle, the behavior of indirect partners is comparable to the behavior of direct partners, which is not surprising as the partners of the partners are dependent on the direct partners. This means that if the number of direct partners of one node is reduced the indirect partners is also likely to be reduced. In general, this means that personal data of users are less likely to be shared unnoticed with multiple parties.

Table 7: Synchronization relations of the top companies observed in our experiments. *Direct Partners* indicates the amount of direct ID syncing. The node types outer (o), balanced (b), center (c), and isolated (iso) are displayed as well.

#	Third Party	Direct Partners				Type			
		M#1	M#2	M#3	M#4	M#1	M#2	M#3	M#4
1.	Google	195	138	118	112	c	c	c	c
2.	Facebook	11	11	9	9	c	c	c	c
3.	Amazon	31	19	17	1	c	c	c	b
4.	Verizon	18	10	10	6	c	c	c	c
5.	AppNexus	69	42	40	44	o	c	c	c
6.	Oracle	30	31	27	18	c	c	c	b
7.	Adobe	11	8	5	4	c	c	b	b
8.	Smart AdServer	1	1	isolated		o	b	isolated	
9.	RTL Group	16	8	7	10	c	c	c	c
10.	Improve Digital	2	1	1	iso	b	b	o	iso
11.	MediaMath	16	7	8	10	c	c	c	c
12.	TripleLift	5	1	2	4	b	o	b	b
13.	RubiconProject	12	isolated			c	isolated		
14.	The Trade Desk	12	7	5	6	c	c	b	c
15.	ShareThrough	2	isolated			b	isolated		
16.	Neustar	isolated			10	isolated			c
17.	Drawbridge	1	iso	1	iso	o	iso	e	iso
18.	Adform	1	14	11	13	o	c	c	c
19.	Bidswitch	3	5	3	2	b	b	b	b
20.	Harris Insights & Analytics	4	2	2	2	b	b	b	b
21.	Acxiom	12	2	6	5	c	b	c	b
22.	Index Exchange	6	4	3	2	c	b	b	b
23.	Criteo	6	1	4	2	c	o	b	b
24.	OpenX	16	7	6	4	c	b	b	e
25.	DataXu	7	6	4	3	b	c	b	b
26.	Lotame	2	3	3	2	o	b	b	b
27.	FreeWheel	isolated				isolated			
28.	Amobee	isolated				isolated			
29.	comScore	25	20	20	17	c	c	c	b
30.	spotX	isolated				isolated			
31.	Sovrn	isolated			2	isolated			b
32.	Sizmek	23	14	18	2	c	c	c	b
33.	Twitter	isolated		2	iso	isolated		o	iso
34.	Microsoft	iso	1	iso	iso	iso	o	iso	iso
35.	Media Innovation Group	2	1	2	2	b	o	o	o
36.	Comcast	5	4	4	4	b	b	b	o
37.	Turn	2	2	2	4	b	b	b	b
38.	Quantcast	2	2	1	1	b	b	o	o
39.	IponWeb	iso	31	isolated		iso	c	isolated	

		Remaining Nodes							
		Mean Direct Partners				Nodes			
	Node Type	M#1	M#2	M#3	M#4	M#1	M#2	M#3	M#4
	Outer corners (o)	1.31	1.26	1.22	1.33	268	183	164	190
	Center nodes (c)	12.56	10	9.33	7.4	25	8	6	5
	Balanced (b)	2.11	2.13	2.09	2.21	205	135	127	106