A Study on Subject Data Access in Online Advertising after the GDPR

Tobias Urban[1][2][0000-0003-0908-0038], Dennis Tatang[2], Martin Degeling[2][0000-0001-7048-781X], Thorsten Holz[2], and Norbert Pohlmann[1]

[1] Institute for Internet Security, Westphalian University of Applied Sciences, Gelsenkirchen, Germany
{lastname}@internet-sicherheit.de

[2] Horst Görtz Institute, Ruhr University Bochum, Bochum, Germany
{firstname.lastname}@rub.de

Abstract. Online tracking has mostly been studied by passively measuring the presence of tracking services on websites (i) without knowing what data these services collect, (ii) the reasons for which specific purposes it is collected, (iii) or if the used practices are disclosed in privacy policies. The European *General Data Protection Regulation* (GDPR) came into effect on May 25, 2018 and introduced new rights for users to access data collected about them.

In this paper, we evaluate how companies respond to *subject access requests* and *portability* to learn more about the data collected by tracking services. More specifically, we exercised our *right to access* with 38 companies that had tracked us online. We observe stark differences between the way requests are handled and what data is disclosed: Only 21 out of 38 companies we inquired (55 %) disclosed information within the required time and only 13 (34 %) companies were able to send us a copy of the data in time. Our work has implications regarding the implementation of privacy law as well as what online tracking companies should do to be more compliant with the new regulation.

Keywords: GDPR, subject access request, privacy, online advertisement

1 Introduction

The business models of modern websites often rely—directly or indirectly—on the collection of personal data. The majority of websites tracks visitors and collects data on their behavior for the purpose of targeted advertising [12]. While in some cases, users knowingly and willingly share personal data, in many other cases, their data is collected without explicit consent or even goes without being noticed [26]. As a result, the imbalance of power over information between data processors (service providers) and data subjects (users) increased in the last years. Furthermore, attackers can also perform a malicious leakage of such data [27].

The *European General Data Protection Regulation* (GDPR) aims to harmonize data protection laws through the EU and to regulate the collection and usage of personal data. Compliance with GDPR is required for any company that offers services in the European Union, regardless of where their headquarter is located. One of the law's goals is to allow users to (re)gain control of the immaterial wealth of their personal data by introducing additional possibilities like the right to request a copy of their data, the right to erasure, and the need for services to explicitly ask for consent before collecting or sharing personal information [14].

Previous work already passively measured the effects of the GDPR. For example, studies analyzed the adoption of privacy policies and cookie consent notices [10,23], while others focused on third parties embedded into websites [6,25].

In this paper, we make use of the new legislation and evaluate the subject access processes of several companies. We identify prominent third parties on popular websites that collect tracking data and exercise our *right to access* with these companies. Besides these two rights, the GDPR also grants the right to erasure, rectification and others that are not part of this work. We provide an in-depth analysis of the processes and show how different companies adopted the new legislation in practice. We analyze timings and success of our inquiries and report on obstacles, returned type of data, and further information provided by companies that help users to understand how personal data is collected. Asides from the detailed overview on different approaches how *subject access requests* (SARs) are implemented in practice; our work provides helpful pointers for companies, privacy advocates, and lawmakers how the GDPR and similar regulation could be improved.

To summarize, our study makes the following contributions:

- We requested access to our personal data from 38 companies and analyze the success of these *subject access requests*. We found that 58 % of the companies did not provide the necessary information within the deadline defined in the GDPR and only a few actually granted access to the collected data.
- We analyzed the privacy policies of these companies regarding usage and sharing of collected personal data. We found that most policies fulfill the minimal requirements of the GDPR, but rarely contain additional information users might be interested in (e. g., partners with whom data is shared).
- Finally, we examined the *subject access request* process of each company and report on data (e. g., clickstream data) and obstacles users face when accessing their data. We found that the provided data is extremely heterogeneous and users sometimes have to provide sensitive information (e. g., copies of identity cards) to access their own data.

2 Background

Our study analyzes the effects of the GDPR as a relatively new legal regulation. We therefore first provide an overview of the GDPR's relevant rules before giving describing the technical background on tracking and data sharing in the online advertisement ecosystem.

2.1 Data Protection Law

The General Data Protection Regulation (GDPR or Regulation 2016/679) [14] is an initiative by the European Union (EU) to harmonize data protection law between its member states. After a transition period of two years, it went 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 data of EU-citizens. Online advertising companies need to disclose, for example, in their privacy policy, for what purpose they collect and share data.

Besides other rights, the GDPR lists the *right to access* (Art. 15) and the *right to data portability* (Art. 20). The difference between those two is that Art. 15 grants users the right to request access to the personal data a company collected about them, while Art. 20 grants users the right to retrieve a copy of the data they provided. According to recital 68 of the GDPR (recitals describe the reasoning behind regulations), the *right to data portability* is meant to support an individual in gaining control over one's personal data by allowing access to the data stored about him or her "*in a structured, commonly used, machine-readable and interoperable format*". For any information request, including those to data access/portability, the GDPR specifies that they must be answered within one month (Art. 12, No. 2), but can be extended by two months.

Some tracking companies claim that the data they use is not personal information because it is anonymized, while in fact is only pseudonymized (see Section 5.1). If the data was anonymous, it would free them from any data protection related obligations, while pseudonymous data that can be attributed to a person using additional information, still falls under the GDPR's rules (Recital 26). In addition, the Article 29 Working Group, a committee of European data protection officials, already made clear in 2010 that storing and accessing a cookie on a user's device is indeed processing of personal data since it "*enable[s] data subjects to be 'singled out', even if their real names are not known*," and therefore requires consent [8]. Relevant for our study is the clarification that ad networks, and not those that embed the third-party scripts on their websites, are responsible for the data processing. Since advertisers rent the space on publisher websites and set cookies linked to their hosts, they are responsible for the data processing, and therefore have to respond to subject access requests.

2.2 Advertising Economy

Displaying ads is the most popular way to fund online services. In 2017, the online advertising industry generated $88.0 billion US dollars [19] in revenue in the US and €41.8 billion Euros in the EU [18]. The ecosystem behind this is complex and is, in a nutshell, composed out of three basic entities which are described in the following [13,28]. On the one end, there are publishers and website owners that use *supply-side platforms* (SSP) to sell ad space (e. g., on websites or prior to videos). On the other end, the *demand-side platform* (DSP) is used by marketing companies to organize advertising campaigns, across a range of publisher. To do so, they 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 digital ad campaigns. They can be used to merge data sets and user information from different sources to automate campaigns on DSPs.

To improve their reach, ad companies utilize *cookie syncing* (sometimes called ID syncing) [22] which allows them to exchange unique user identifiers. Using this method, companies can share information on specific users (e. g., sites on which they tracked them) and learn more about the user. 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). RTB involves that impressions and online ad space are sold in real-time on automated online marketplaces whenever a website is loaded in a browser.

3 Related Work

Most previous work analyzes online privacy through measurements (e. g., [12, 20]), but these studies have all been conducted prior to the GDPR. With the introduction of the GDPR, several research groups started measuring the effects of the legislation. Degeling et al. analyzed the adoption and effect of the GDPR regarding privacy policies and cookie notice banners [10]. 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 [7]. In contrast, Sørensen et al. found that the number of third parties did slightly decline since the GDPR went into effect, but they conclude that the GDPR is not necessarily responsible for that effect [25]. Boniface et al. analyze the tension between authentication and security when users perform a SAR [4] and discuss measures used to identify users and discuss threats (e. g., denial of access) of too harsh measures. In line with our findings, they also report on disproportional identity checks. Most recently, two studies analyzed how adversaries could abuse subject access requests to get access to personal data of other individuals [5,11]. Both studies spoof an identity and request access to personal data of the spoofed identity and by this they show that SARs are often not adequately verified and therefore, companies unintentionally leak personal data. De Hert et al. [17] discuss the right to data portability from a computer law point of view. They give a systematic interpretation of the new right and propose two approaches on how to interpret the legal term “data provided” in the GDPR. The authors argue that a minimal approach, where only data are directly given to the controller 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”.

4 Study Design

To gain insights into the way how companies grant access to collected personal data, we first identified prominent companies often embedded into websites, and afterward, we exercised our right to access/portability with these companies.

4.1 Approach

Our study consist of two steps: First, we *passively* measure the most prominent companies used as third parties on websites and the companies most active in sharing personal identifiers. Afterward, we *actively* measure and analyze the information provided by companies to users that file a subject access request.

In order to identify the most prominent companies, we perform a three-step measurement (see Sec. 4.2): (1) We visit a number of websites, (2) extract the third parties embedded in these websites, and (3) extract all ID syncing activities from the observed requests. Based on the gathered information, we determine top companies that engage in ID syncing and top companies that are often embedded into websites. We did choose to focus on top embedded companies as these potentially affect most users and more users might issue a *subject access request* (SAR) to these companies. Furthermore, we choose the top syncing parties as these might share personal data of users without properly informing them—which would make it quite hard for users to actually regain control of their personal data if they do not know who holds their data. In order to learn more about the privacy practices from the companies themselves, we analyze privacy policies to see if the data sharing and other necessary information are made transparent to users (see Sec. 4.3). Then we use our right to access/portability to learn how companies respond to SARs and which data they provide to users.

In our experiments, we use the *openWPM* [12] platform and deployed it on two computers located at a European university. Thus, our traffic originates in the EU (and ultimately from an EU resident who started the crawl) and therefore the GDPR applies. *OpenWPM* was configured to log all HTTP request and response with the corresponding HTTP 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 did allow third-party cookies.

To analyze the sharing of *digital identifiers* (IDs), we first have to define them. For every visited domain, we analyzed the HTTP `GET` and `POST` requests and split the requests at characters that are typically used as delimiters (e. g., ‘&’ or ‘;’). As a result, we obtained a set of ID candidates that we stored as key-value pairs for later analysis. We identified IDs according to the rules previously defined by Acar et al. [1] (e. g., IDs have to be of a certain length or must be unique). To measure the syncing relations of third parties, it is necessary to identify URLs—that contain user IDs—inside a request (e. g., **foo.com/sync?partner=https://bar.com?/id=abcd-1234**). According to the named rules, we parsed all URLs and checked if an HTTP parameter contains an ID, Furthermore, we used the *WhoTracks.me* database [6] to cluster all observed third-party websites based on the company owning the domain.

4.2 Analysis Corpus

The complexity of the online advertising ecosystem was already highlighted in previous work [3, 16]. To the best of our knowledge, there is no reliable public information on market shares in the online advertising ecosystem. Thus, we performed an empirical measurement and identified the top companies in that measurement. To identify the most popular companies, we visited the Alexa top 500 list [2] and randomly visited three to five subsites of each domain. We visited the selected websites using the *openWPM* setup described above.

We selected the 25 most embedded third parties as well as the top 25 third parties that engaged most in cookie syncing for in-depth analysis of what information they share with users. In total, we identified 36 different companies which we refer to as *analysis corpus*. In three cases, we were told during the SAR process to address our inquiry to another company so that our final corpus consists of 39 companies. In the remainder, if not stated otherwise, our analyses of privacy policies and information disclosure refer to this corpus.

The first company that we did *not* include in the corpus (i. e., the 26th most embedded company) was embedded by just 0.12 % of the visited websites and the first ID syncing company *not* included in the corpus accounts for 0.58 % of the syncing connections in the graph. The 39 companies in the corpus account for 66 % of all ID syncing activities, while the reaming 33 % are made up of 352 companies. The companies in the corpus represent 61 % of the embedded third parties. Contacting ten more companies (an increase of 19 %) would increase the amount of covered ID syncing by at most 5.8 % or embedded websites by at most 1.2 %. The corpus consists of six SSPs, nine DSPs, seven companies that specialized in targeted ads, four DMPs, and 13 companies whose primary business field is not directly tied to the advertising but instead utilizes ads to finance their services (e. g., *RTL Group*—a Luxembourg-based digital media group).

While most of the companies in our corpus operate globally and run multiple offices, 82 % have their headquarters located in the United States. The remaining 18 % are located in Europe. This distribution is likely based towards US/EU-based companies since we run our measurements from Europe. We discuss the limitations of our analysis corpus in Section 6.

4.3 Transparency Requirements

The privacy policies of all 39 companies described above were analyzed by a certified data protection expert with a computer science background to see whether they contain the information required by the GDPR (see Sec. 2). We specifically looked for information on data sharing practices and evaluated how data subjects can exercise their rights. As described above, data controllers are required to inform, besides other things, about the legal basis for their data collection, categories of companies they share the data with, and how long the data is stored. We do not report on observations that all policies had in common but focus on the differences. On the one hand, for example, the right to withdraw consent has been implemented through various opt-out mechanisms [10] that all services support and are therefore not listed. On the other hand, few services actually follow the "Do Not Track" signal, although it was designed as a common consent mechanism. Therefore, we listed statements about the latter. We were also interested in how companies deal with the requirements regarding profiling: If they use profiling, they are expected to describe *the logic involved* in this process, although the debate about what that should include is still ongoing [24]. Privacy policies should list the rights of the data subjects, e. g., to object to the processing and the possibility to access the data and they should describe how these rights can be exercised. While the policies should also specify whether data is shared with third parties, companies are not required to list them individually but can describe them in categories.

4.4 Assessing the SAR Process

In order to test to which extent users can actually exercise data access rights, we reached out to companies in the corpus after extracting contact information from their privacy policies. According to Articles 13 and 14 GDPR, contact details of a responsible person (e. g., the Data Privacy Officer) need to be provided for privacy-related questions. Most companies (27) named a general email address to handle such requests or referenced a web form to access the data.

In our requests, we referenced a profile that was generated specifically for this process. We used *openWPM* to randomly visits websites that include third parties, owned by the companies in our analysis corpus. From these websites, all internal links (subsites) were extracted and visited in random order. For this analysis, we kept the session active and continued visiting websites while we requested information about the profiles. This *openWPM* instance was left running until the end of our analysis in order to keep the cookies active.

When sending out inquiries, we included all cookie IDs and domains for which we observed ID syncing (with the corresponding IDs). If we could add custom text to our request (via email and in some web forms), we asked four questions regarding the usage of our data: (1) *What information about me/associated with that cookie do you store and process?*, (2) *Where did you get that information from? Did you get it from third parties?*, (3) *Do you use the data to perform profiling?*, and (4) *With whom do you share what information and how?*

We used informal language (e. g., we did not quote any articles from the GDPR nor did we use any legal terminology) because we wanted to assess the process when users with some technical understanding of online advertising (e. g., users who can read cookies from the cookie store), but no legal background, want to exercise their right to access/portability. Actual users might have trouble to access the information we added in our emails (e. g., the correct cookie values). However, some companies offer simplified ways to access the information to be included in requests (e. g., a web form that reads the user ID from the browser's cookie store). We assume that a user who has privacy concerns can obtain this information and usability improvements might follow in the near future.

We conducted two rounds of inquiries. The first round in June 2018, one month after the GDPR took effect, and the second round was starting three months later, in September 2018. We did so to make sure answers were not biased by being the first ones the companies received. We used two *GMail* accounts we created for this purpose (one for each round of contacting) to get in touch with the companies and did *not* disclose that we were conduction this survey to avoid biased responses. The response timings were evaluated in relation to two deadlines: The first deadline is the legal period defined in the GDPR, 30 days after the request, and a more relaxed deadline 30 *business* days after our requests.

5 Results and Evaluation

Companies are required to share certain information, e. g., who has access to the data and where it is transferred to, publicly in their privacy policies. Other information, e. g., what is stored about a user, has to be disclosed upon request.

5.1 Evaluation of Privacy Policies

We analyzed the privacy policies of the companies in our corpus to check whether they fulfill the requirements described in Section 4.3. The most relevant details are reported in Table 1 (see Appendix B). All but three policies fulfill the minimum requirements for privacy policies set by the GDPR, all companies offer the possibility to opt-out of their services, and all except one disclose that they share some information with others. At the same time, only three are transparent about who these third parties are and what type of information is actually shared. Only two of the policies disclosed and explain cookie syncing. Similarly, only eight mention whether or not they perform profiling. One company did not update its privacy policy since 2011 and it contained false claims, for example, that IP addresses are non-personal information. *Amazon*'s privacy policy was least transparent concerning the information we were looking for.

All policies except for four policies mention a legal basis for processing, which is now required. 31 claim that they rely on individual consent when processing data but at the same time only three mention that they adhere to the “Do Not Track” (DNT) standard, where information about whether or not users want to be tracked is conveyed in an HTTP header [21]. Instead, companies refer to implicit consent, which implies consent as long as a data subject has not manually objected to a data collection by opting-out.

Differences can be found on topics specific to GDPR, for example, regarding the question of whether a company processes data that contains sensitive information (e. g., about race or health). While 13 explicitly forbid to collect this information through their services, four acknowledge that some interest segments they provide might be health-related e. g., about beauty products. Three companies acknowledge that they process health-related information, but do not discuss how this data is better protected than the rest. The majority (17) does not make any statements about their practices in this area.

5.2 Subject Access Requests

In order to analyze the process how users can access personal data collected about them and to fill the blanks left by the privacy policies, we examined how third parties adopt the new requirements of the GDPR (see Section 2.1) and how they respond to *subject access requests* (SAR).

We contacted the companies in our analysis corpus and tried to exercise our right to access and right to portability of the data associated with a cookie ID to evaluate the SAR process of each company as described in Section 4.4. In the first round (June 20th, 2018), we sent out 32 emails and used six web forms to get in touch with each company. In the second round (September 21th, 2018), we sent 27 emails and used eleven web forms as the contact mechanisms had slightly changed. As part of this process, we extracted the cookie ID values and up to five domains associated with each company for which we observed ID syncing (with the ID key-value pairs) from the long-running profile in the email. The GDPR requires companies to grant users access to their data within 30 days after their initial request. Since it does not specify whether these referrers to business or calendar days, we marked two deadlines (dotted, gray lines in Figs. 1 and 2).

Response Types and Timing We grouped responses in three types: (1) *automatic* responses, (2) *mixed* responses, and (3) *human* responses. A message was categorized as "automatic" if it was identifiable as sent automatically by a computer system (e. g., a message from a ticket system). We labeled a message *mixed* if the message did not directly refer to any of our questions but only included very generic information that responds to any privacy-related request. Messages that directly responded to our questions were labeled "human". To increase the accuracy of the classification, we compared the content from both inquiry rounds and if there was any doubt, we ruled in favor of the companies. Figure 1 shows amounts and type of responses we got during our analysis. We did not count status messages from ticket systems (e. g., a message stating that our email was received) but only looked at those messages that contained an actual reply.

In our second round of inquiries, we received fewer responses (approx half of the amount). This is partly because we did not have to report any broken data access forms, that we encountered in round one, to companies which explain the fewer human responses in weeks one and two. However, we observed that in our second round, companies did not follow up further questions as they did in round one (e. g., if we asked for further clarification about data sharing).

In round one, we received the most responses (51/100) during the first two weeks, where we labeled the majority as send by a *human* (57 %) and 26 % as *automatic*. While the share of response types stayed balanced, the number of responses significantly decreased (by 43 %) in the following weeks, although we asked follow-up questions. In round two, these types of answers changed as we considered 17 % of the responses as sent by a *human* and 61 % *automatic*.

While we still received responses from human correspondents one week before the deadline in the first round, responses were lower in the second round (a third).

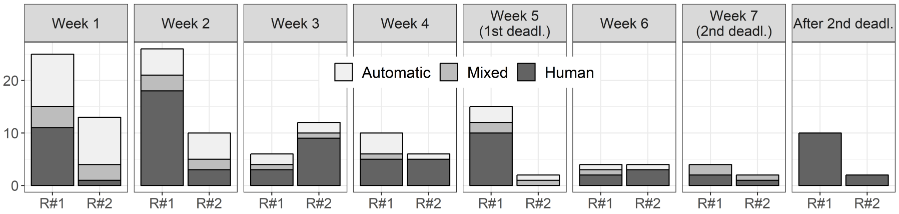


Fig. 1: Types and timings of the received responses.

Only one company told us (in both rounds) that due to the complexity of our inquiry that they would need more time.

Response Success The effort necessary to obtain access to personal data differed depending on the inquired company. To asses the workload of the process of a company, we use a simple scoring mechanism that essentially takes four factors with different impact into account: (1) amount of emails sent to the company before getting access to data associated to the digital ID, (2) amount of emails sent after getting access to the data, (3) actions that the user has to perform online, and (4) actions that a user has to perform offline. These simple metrics do not account for the actual effort each obstacle might pose to an individual asking for access, but it is helpful to approximate the complexity of the process.

We differentiate between emails because we interpret access to collected data as the primary goal of the request. However, there might still be some open questions (e. g., if profiling is performed) that were not answered by the time the data was shared. An example of an action that a user must perform online is that the user has to enter additional data in an online form (e. g., legal name). On the contrary, scanning the user's official identification document (e. g., passport) is a typical example of a task a user has to perform offline. We created our "workload score" to measure (1) if companies set up obstacles, (2) if companies ask for additional information, and (3) the amount of interaction necessary.

In Section 5.2, we describe the procedure of how users can access personal data (of the companies in our analysis corpus) in more detail. The result of the workload determination and comparison between inquired companies is given in Figure 2. The figure shows a clustered version of the SAR results. We computed the distance between all points of the same "response status" (e. g., "got access") and clustered the points that are close to each other. The larger and higher each point, the more companies asked for more effort to answer our requests.

Table 3b shows the results of our inquiries by the time of the first deadlines (July 20th/October 31st). Note that is unlikely that we provided a wrong cookie ID, but it is possible that a company does not have any data on the record because of short retention times or when some events are not logged, because there was no further interaction. It is notable that some companies stated that if one does not have a user account on their website, they will not store any data related to a cookie ID. These companies did not respond to our SAR request within the legal deadline, in round two. One of these companies replied with our second deadline stating that they do not store any data related to the cookie ID.

Eight companies interpreted the start date of the process as the day on which they got all the administrative data they need to process the inquiry. In all cases, it was virtually impossible for users to know upfront that this data was needed since the companies only shared the needed documents via email and did not mention them in their privacy policies (e. g., one company replied after seven days and asked for a signed affidavit. After we provided the affidavit, they told us, five days later, that they would *"start the process"* and reply within 30 days.).

In total (after the second deadline of round one), only 21 of 36 companies (54 %) shared data, or told that they do not store any data, 15 of 36 (42 %) were still in the process (or did not respond), and one company said that it would not share the data with us because they cannot properly identify us. In round two, 64 % granted access or told us that they do not store any data, 33 % did not finish the process, and again one company declined to grant access since they could not identify us. In these numbers, we *excluded* companies that told us to address a subsidiary/parent company with our inquiry.

Figure 2 shows that if companies granted access, we see that the workload is often quite low (in both rounds). In one case with high workload, in round one, a long email exchange (in total 13 emails—six sent by us) was needed to get access, the other cases required a copy of the ID and in one case a signed affidavit. It is notable that the overall workload in round 2 lowered and companies usually wrapped up the process faster. The reduction of workload is because, on the one hand, we did not have to report broken SAR forms and on the other hand companies set up less "offline" obstacles.

Especially during round one, we observed that companies who claimed not to store any data still require multiple interactions prior to providing that information. Two companies required a signed affidavit and a photocopy of an ID. The third company, after a long email conversation, asked to call the customer support to explain our case in more detail, still coming to the result that they do not store any data. All three companies did not respond in round two.

Disclosed Information Figure 3c gives an overview of the data we received as a result of the SARs. We categorized the received data in terms of readability and content. If data was presented in a way a human can easily read it (e. g., on a website), we labeled it "*human readable*" and otherwise "*raw*" (e. g., `.csv` files). If the data contained visited websites, we labeled it "*Tracking*", if it contained segment information, associated with the profile, we labeled it "*Segment*", and if it contained the location of the user, based on the used IP address, we labeled it "*Location*". Otherwise, we labeled it "*Other*".

The shared data was heterogeneous in format (e. g., `.pdf`, `.csv`, `.htm`, etc.), data contained (e. g., interest segments, clickstream data, IP addresses, etc.), and explanation of the data (examples of shared data are provided in Appendix A. One company shared an `.csv` file with headers named c_1 to c_{36} (sic.), while another company provided detailed explanations in an appended document and yet another told us that we should contact them if we had trouble understanding the

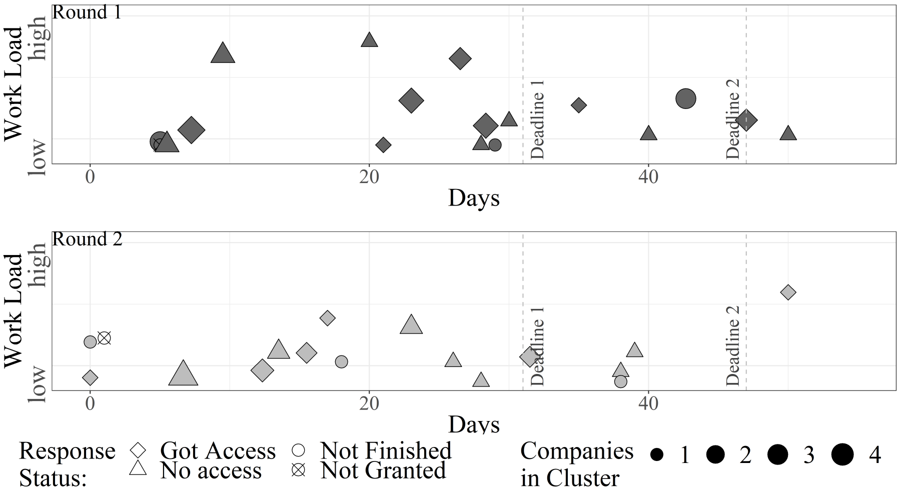


Fig. 2: Comparison of the workload to get access to personal data companies stored about a user.

data. If a company shared clickstream data (three in total), we manually checked if the data set contained additional or missing websites that we had observed. In all three cases, the data was accurate. Overall, the received data can be grouped into three categories: (1) technical data, (2) tracking data, and (3) segment data. *Technical data* is raw data, often presented in text files, the companies directly extracted from HTTP traffic (see Fig. 5). *Tracking data* is information on which websites the company has tracked the user, also typically presented in a text file (see Fig. 6). *Segment data* is data companies inferred from a user's online behavior (see Fig. 4), which was typically presented on a website (e. g., user interests). In terms of clarity of the provided data, we also found different approaches. Some companies shared segments they inferred from our (artificially) browsing behavior (e. g., Segment: *Parenting - Millennial Mom* (sic.)), others shared cryptic strings without explanation (e. g., *Company-Usersync-Global*), or data that was incorrectly formatted somewhere in the process to the point where it was almost unintelligible (e. g., *Your_hashed_IP_address: Ubuntu* (sic.)). However, we did not find any instance where data was provided that was not mentioned to be collected in the privacy policy and many instances (all but one) where not all data that might be collected was provided.

Subject Access Request Process Companies handle inquiries very differently ranging from not responding at all, over simply sending the personal data via email, to sending (physical) letters which had to include a copy of a government-issued identification card and a signed affidavit, stating that the cookie and device belong to the recipient and only the recipient.

Fig. 3: Overview of the SAR process and responses for both rounds of inquiries.

(a) Obstacles

Status	R1	R2
Affidavit	4	3
ID card	6	5
Other	4	7
None	26	25

(b) Response success

Status	R1		R2	
Access	14	39 %	8	22 %
No Data	7	19 %	13	36 %
Denied	1	3 %	1	3 %
Not Finished	11	31 %	9	25 %
No Response	4	11 %	5	14 %

(c) Response data

Type	R1	R2
Raw data	9	3
Human read.	5	5
Segments	4	4
Tracking	3	3
Location	4	4
Others	5	2

(d) Answers

Question	R1	R2
Q1 (data)	21	23
Q2 (sources)	6	6
Q3 (profiling)	9	6
Q4 (sharing)	7	4

Table 3a gives an overview of the obstacles users face when filling a SAR. Most companies require the user to provide the digital identifier (or directly read it from the browser's cookie storage) in order to grant access to the data associated with it. Since most online forms do not provide all data, a company collected about the user (e. g., they provide the ad segments associated with the user but not the used IP addresses or visited websites) it is reasonable to grant access to this data if the cookie ID is provided. However, online forms come with the risk that an adversary might fake the cookie ID to get access to personal data that is associated with another individual. An affidavit is a way to counter this sort of misuse, and one company stated this as the reason for this step.

The GDPR states companies "*should use all reasonable measures to verify the identity of a data subject who requests access*", to make sure they do not disclose data to the wrong person. Asking for identifying information is supposed to add a layer of security when data subjects request a copy of their data. The ad industry association emphasizes the possibility of this additional safeguard [15], but official interpretations state that data processors should have "*reasonable doubts*" before asking for additional data [9]. Those that request an ID card did not explain their doubt and did not describe how the ID helps them to verify that the person requesting the data actually owns the cookie ID.

Answers to Our Questions Finally, we want to discuss the answers to the four questions we asked in the inquiries (see Sec. 5.2). Only a few companies did answer the additional questions we asked. Most of them referred to their privacy policy or did not provide further details. Table 3d gives an overview of the responses we got to our questions. Note that companies were not obliged to answer the question and that we could not check if they answered truthfully—if there is no public information in e. g., the privacy statements that say otherwise (see Sec. 6). With respect to Q1 and Q2, most answers contained references to or parts of the privacy policy. As Table 1 (Appendix B) shows, only a few companies (nine/seven) disclose whether or not they perform profiling. Only one of the answers, where the privacy policy was unspecific, clearly stated that the data is not used for profiling. Six answers described in more detail how the data is processed and would suffice the GDPR rule that "*meaningful information* *about the logic involved*" should be provided. One company stated in their email that they do not perform profiling, although their privacy policy mentions it. Unfortunately, only seven/five companies listed their actual sharing partners. When companies stated with whom specifically they shared our data (i. e., not a general list of partners), we could confirm this through our measurement, but in three cases companies stated that they did share data with specific companies that were not listed in their privacy policy. The low amount of companies that named partners with whom they share data poses a problem for users that want to understand who has received a copy of their personal information.

6 Limitations and Ethical Considerations

We contacted 39 companies, which represents only a small subset of all online advertising companies. However, we showed that the contacted companies come from different market areas and that they represent the most prominent companies (in our measurement). Future work should focus on the usability of SARs in a user study and include more companies. Similarly, our scale to visualize the complexity of the subject access requests (Fig. 2), should be validated with user experiments. Right now it serves only as an approximation.

Since our research includes human subjects (the persons exercising their rights and the persons responding to our requests), ethical considerations need to be taken into account. In this work, we analyze the SAR process of different companies and not the *persons* replying in detail. Hence, we do not see any particular reason why we have to disclose that we conduct this survey. Note that after our second deadline (in our first measurement), we contacted the companies that did not respond at all or had a poorly designed process, without any responses. When contacting the companies, we did not disclose we conduct a scientific survey, but we did disclose the real names of two authors in each mail and on the photocopied IDs. We also answered all of the companies questions truthfully (e. g., if we had been in contact with a company in any other way aside from this survey) and reported all problems (e. g., broken data access forms) that we noticed during the process.

7 Conclusion

Our work shows that while most companies offer easy ways to access the collected personal data, few disclose all the information they have and some companies create significant obstacles for users to access it. The obstacles range from signed affidavits over providing additional information (e. g., phone numbers) to copies of official ID documents. Some larger companies do not disclose data to users that are not registered with their services. The different approaches of how access to personal data is granted show the different interpretations of the new law. Looking into the response behavior, we see that over 58 % of the companies did not respond within the legal period of 30 days, but only one company extended the deadline by two more months.

Acknowledgment

This work was partially supported by the Ministry of Culture and Science of the German 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.

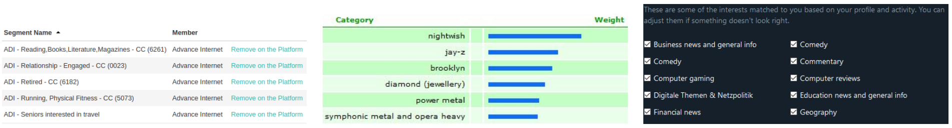
References

1. Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., Diaz, C.: The Web Never Forgets. In: Proceedings of the 21st ACM Conference on Computer and Communications Security. pp. 674–689. CCS'14, ACM Press (2014)
2. Alexa: Top sites for countries (2018), `https://www.alexa.com/topsites/countries`, accessed: 2019-02-05
3. Barford, P., Canadi, I., Krushevskaja, D., Ma, Q., Muthukrishnan, S.: Adscape: Harvesting and analyzing online display ads. In: Proceedings of the 23rd World Wide Web Conference. pp. 597–608. WWW'14, ACM Press (2014)
4. Boniface, C., Fouad, I., Bielova, N., Lauradoux, C., Santos, C.: Security Analysis of Subject Access Request Procedures How to authenticate data subjects safely when they request for their data. In: 2019 - Annual Privacy Forum. pp. 1–20 (2019)
5. Cagnazzo, M., Holz, T., Pohlmann, N.: Gdpirated–stealing personal information on- and offine. In: Proceedings of the 2019 European Symposium on Research in Computer Security. ESORICS'19, Springer-Verlag (2019)
6. Cliqz: Whotracks.me data - tracker database (2018), `https://whotracks.me/blog/gdpr-what-happened.html`, accessed: 2019-04-24
7. Dabrowski, A., Merzdovnik, G., Ullrich, J., Sendera, G., Weippl, E.: 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 (2019)
8. Data Protection Working Party: Opinion 2/2010 on online behavioural advertising (2010)
9. Data Protection Working Party: Article 29—guidelines on the right to data portability. Tech. Rep. 16 /EN WP 242, European Commission (Dec 2016)
10. Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., Holz, T.: 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 (2019)
11. Di Martino, M., Robyns, P., Weyts, W., Quax, P., Lamotte, W.L., Andries, K.: Personal information leakage by abusing the GDPR ”right of access”. In: Proceedings of the 15th Symposium on Usable Privacy and Security. SOUPS'19, USENIX Association (2019)
12. Englehardt, S., Narayanan, A.: Online tracking: A 1-million-site measurement and analysis. In: Proceedings of the 2016 ACM Conference on Computer and Communications Security. pp. 1388–1401. CCS'16, ACM Press (2016)
13. Estrada-Jimnez, J., Parra-Arnau, J., Rodrguez-Hoyos, A., Forn, J.: Online advertising: Analysis of privacy threats and protection approaches. Computer Communications **100**, 32–51 (2017)
14. European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council (2016), `http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC`

15. GDPR Implementation Working Group: Data subject requests. Tech. Rep. Working Paper 04/2018 v1.0, IAB Europe (Apr 2018), https://www.iabeurope.eu/wp-content/uploads/2018/04/20180406-IABEU-GIG-Working-Paper04_Data-Subject-Requests.pdf
16. Guha, S., Cheng, B., Francis, P.: Challenges in measuring online advertising systems. In: Proceedings of the 10th Internet Measurement Conference. pp. 81–87. IMC'10, ACM Press (2010)
17. Hert, P.D., Papakonstantinou, V., Malgieri, G., Beslay, L., Sanchez, I.: The right to data portability in the gdpr: Towards user-centric interoperability of digital services. Computer Law & Security Review **34**(2), 193–203 (2018)
18. IAB Europe: European digital advertising market has doubled in size in 5 years (2017), https://www.iabeurope.eu/research-thought-leadership/resources/iab-europe-report-adex-benchmark-2017-report/. Accessed: 2019-02-05
19. Interactive Advertising Bureau: Internet advertising revenue report (2017), https://www.iab.com/wp-content/uploads/2018/05/IAB-2017-Full-Year-Internet-Advertising-Revenue-Report.REV2_.pdf. Accessed: 2019-04-24
20. Karaj, A., Macbeth, S., Berson, R., Pujol, J.M.: Whotracks.me: Monitoring the online tracking landscape at scale. CoRR **abs/1804.08959** (2018)
21. McDonald, A., Peha, J.M.: Track Gap: Policy Implications of User Expectations for the 'Do Not Track' Internet Privacy Feature. SSRN Scholarly Paper, Social Science Research Network, Rochester, NY (2011)
22. Papadopoulos, P., Kourtellis, N., Markatos, E.P.: The Cost of Digital Advertisement. In: Proceedings of the 2018 World Wide Web Conference. pp. 1479–1489. WWW'18, International World Wide Web Conference Committee (2018)
23. Sanchez-Rola, I., Dell'Amico, M., Kotzias, P., Balzarotti, D., Bilge, L., Vervier, P.A., Santos, I.: 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. pp. 340–351. ACM Press (2019)
24. Selbst, A.D., Powles, J.: Meaningful information and the right to explanation. International Data Privacy Law **7**(4), 233–242 (Nov 2017)
25. Sørensen, J.K., Kosta, S.: 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 Steering Committee (2019)
26. TRUSTe, Interactive, H.: Consumer research results - privacy and online behavioral advertising (2011), https://www.eff.org/files/truste-2011-consumer-behavioral-advertising-survey-results.pdf. Accessed: 2019-04-24
27. Urban, T., Tatang, D., Holz, T., Pohlmann, N.: Towards understanding privacy implications of adware and potentially unwanted programs. In: Proceedings of the 2018 European Symposium on Research in Computer Security. pp. 449–469. ESORICS'18, Springer-Verlag (2018)
28. Yuan, Y., Wang, F., Li, J., Qin, R.: A survey on real time bidding advertising. In: Proceedings of the 2014 Conference on Service Operations and Logistics, and Informatics. pp. 418–423. SOLI'14, IEEE (2014)

A Provided data

In this section, we provide examples of different types of data we received when performing the subject access requests. The data can be grouped into three categories: (1) "interest segments"—inferred from the user's online activities (Fig. 4), (2) "technical data"—extracted from HTTP traffic (Fig. 5), and (3) "tracking data"—websites on which users were tracked (Fig. 6).



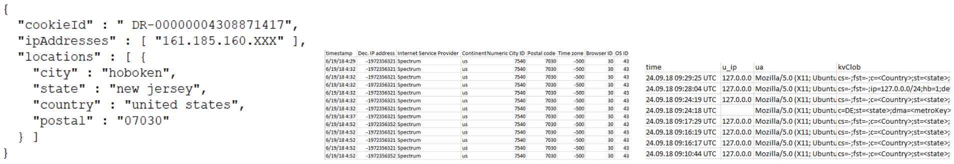
Segment Name ▲	Member	
ADI - Reading,Books,Literature,Magazines - CC (6261)	Advance Internet	Remove on the Platform
ADI - Relationship - Engaged - CC (0023)	Advance Internet	Remove on the Platform
ADI - Retired - CC (6182)	Advance Internet	Remove on the Platform
ADI - Running, Physical Fitness - CC (5073)	Advance Internet	Remove on the Platform
ADI - Seniors interested in travel	Advance Internet	Remove on the Platform

Category	Weight
nightwish	
jay-z	
brooklyn	
diamond (jewellery)	
power metal	
symphonic metal and opera heavy	

These are some of the interests matched to you based on your profile and activity. You can adjust them if something doesn't look right.

☑ Business news and general info
☑ Comedy
☑ Computer gaming
☑ Digitale Themen & Netzpolitik
☑ Financial news
☑ Comedy
☑ Commentary
☑ Computer reviews
☑ Education news and general info
☑ Geography

Fig. 4: Inferred *interest segments* provided by different companies (anonymized).

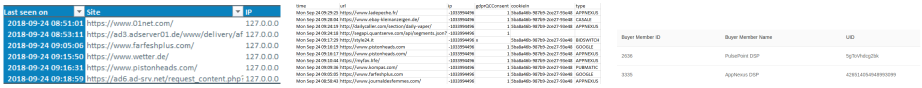


```
{
  "cookieId" : " DR-00000004308871417",
  "ipAddresses" : [ "161.185.160.XXX" ],
  "locations" : [ {
    "city" : "hoboken",
    "state" : "new jersey",
    "country" : "united states",
    "postal" : "07030"
  } ]
}
```

timestamp	Dec. IP address	Internet Service Provider	Continent	Numeric City ID	Postal code	Time zone	Browser ID	OS ID
6/19/18 4:29	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:32	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:32	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:32	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:32	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:32	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:32	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:37	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:52	-1972356352	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:52	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:52	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:52	-1972356321	Spectrum	us	7540	7030	-500	30	43
6/19/18 4:52	-1972356352	Spectrum	us	7540	7030	-500	30	43

time	u_ip	ua	kvClob
24.09.18 09:29:25 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu	cs=-;fst=-;c=<Country>;st=<state>;
24.09.18 09:28:04 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu	cs=-;fst=-;ip=127.0.0.0/24;hb=1;de
24.09.18 09:24:19 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu	cs=-;fst=-;c=<Country>;st=<state>;
24.09.18 09:24:18 UTC		Mozilla/5.0 (X11; Ubuntu	c=DE;st=<state>;dma=<metroKey>
24.09.18 09:17:29 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu	cs=-;fst=-;c=<Country>;st=<state>;
24.09.18 09:16:19 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu	cs=-;fst=-;c=<Country>;st=<state>;
24.09.18 09:16:17 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu	cs=-;fst=-;c=<Country>;st=<state>;
24.09.18 09:10:44 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu	cs=-;fst=-;c=<Country>;st=<state>;

Fig. 5: *Technical data* provided by different companies (anonymized).



Last seen on	Site	IP
2018-09-24 08:51:01	https://www.01net.com/	127.0.0.0
2018-09-24 08:53:11	https://ad3.adserver01.de/www/delivery/af	127.0.0.0
2018-09-24 09:05:06	https://www.farfeshplus.com/	127.0.0.0
2018-09-24 09:15:50	https://www.wetter.de/	127.0.0.0
2018-09-24 09:16:31	https://www.pistonheads.com/	127.0.0.0
2018-09-24 09:18:59	https://ad6.ad-srv.net/request_content.php?	127.0.0.0

time	url	ip	gdprQCConsent	cookieIn	type
Mon Sep 24 09:29:25	https://www.ladepeche.fr/	-1033994496	1	5ba8a46b-987b9-2ce27-93e48	APPNEXUS
Mon Sep 24 09:28:04	https://www.ebay-kleinanzeigen.de/	-1033994496	1	5ba8a46b-987b9-2ce27-93e48	CASALE
Mon Sep 24 09:24:19	http://dailycaller.com/section/daily-vaper/	-1033994496	1	5ba8a46b-987b9-2ce27-93e48	APPNEXUS
Mon Sep 24 09:24:18	http://segapi.quantserve.com/api/segments.json?	-1033994496	1		
Mon Sep 24 09:17:29	http://style24.it	-1033994496	x	5ba8a46b-987b9-2ce27-93e48	BIDSWITCH
Mon Sep 24 09:16:19	https://www.pistonheads.com	-1033994496	1	5ba8a46b-987b9-2ce27-93e48	GOOGLE
Mon Sep 24 09:16:17	https://www.pistonheads.com/	-1033994496	1	5ba8a46b-987b9-2ce27-93e48	APPNEXUS
Mon Sep 24 09:10:44	https://myfav.life/	-1033994496	1	5ba8a46b-987b9-2ce27-93e48	APPNEXUS
Mon Sep 24 09:09:36	https://www.kompas.com/	-1033994496	1	5ba8a46b-987b9-2ce27-93e48	PUBMATIC
Mon Sep 24 09:05:05	https://www.farfeshplus.com	-1033994496	1	5ba8a46b-987b9-2ce27-93e48	GOOGLE
Mon Sep 24 08:58:43	https://www.journaldesfemmes.com/	-1033994496	1	5ba8a46b-987b9-2ce27-93e48	APPNEXUS

Buyer Member ID	Buyer Member Name	UID
2636	PulsePoint DSP	5gToVhdcg2bk
3335	AppNexus DSP	426514054948993099

Fig. 6: *Tracking data* provided by different companies (anonymized).

B Privacy Policy Overview

Table 1 provides a summary of the privacy policies of the companies in our data set. It lists the most important tracking and GDPR-related attributes and what information is disclosed.

Table 1: Overview of information available in privacy policies. * marks information that is required by the GDPR. *Legal Basis* refers to the sections in Article 6 of the GDPR: (a) consent, (b) contract, (c) legal obligation, (e) public, (f) legitimate interest; n.m. = not mentioned

Company	Legal Basis*	Shared Data	3rd CO*	Sensitive Data	Profiling	Retention*	Partners*	Data Access*	DNT	Version
Google	a,b,c,f	unspecified	y	n.m.	n.m	unspecified	7	account	n.m.	05/2018
Facebook	a,b,c,d,e,f	unspecified	y	y	n.m.	differs	categories	account	n.m.	04/2018
Amazon	n.m.	unspecified	n.m.	n.m.	n.m.	n.m.	categories	n.m.	n.m.	08/2017
Verizon	a,b,c,f	unspecified	y	n.m.	n.m	unspecified	329	website, email	n.m.	05/2018
AppNexus	a,f	unspecified	y	n.m.	n.m	3-60d, up to 18m	2309	website	n.m.	05/2018
Oracle	a,c,f	unspecified	y	health related	n.m.	12-18m	categories	website	y	05/2018
Adobe	a,b,c,f	unspecified	y	n.m.	y	until opt-out	categories	email, form	n	05/2018
Smart AdServer	a,f	unspecified	y	n.m.	y	1d-13m	categories	email	n.m.	05/2018
RTL Group	a,c,f	unspecified	y	n.m.	n.m.	as long as necessary	categories	email	n.m.	unclear
Improve Digital	a	listed	y	n.m.	n.m.	90d	categories	email	y	05/2018
MediaMath	f	unspecified	y	health related	y	up to 13m	categories	email	n.m.	05/2018
Triplelift	a,f	unspecified	y	ask to avoid	n.m	as long as necessary	categories	website	n	05/2018
RubiconProject	a,b,c,f	unspecified	y	n.m.	n.m.	90-366d	categories	form	n.m.	05/2018
The Trade Desk	a,f	unspecified	US	not allowed	n.m.	18m, 3y aggregated	categories	website	n.m.	10/2018
ShareThrough	a,b,c,f	unspecified	y	n.m.	y	13m	categories	email	n.m.	05/2018
Neustar	n.m.	IDs, segments	US	not allowed	n	13m + 18m aggregated	categories	email	n.m.	08/2018
Drawbridge	n.m.	IDs, segments	US	health related	n.m.	n.m.	categories	email	n	08/2018
Adform	a,f	unspecified	y	not allowed	n.m.	13m	33	form/email	n.m.	unclear
Bidswitch	a,b,c,f	unspecified	y	n.m.	n.m.	“as long as necessary”	categories	n.m.	n.m.	05/2018
Harris I & A	a,c	listed	y	y	n.m.	purpose fulfilled	categories	email	n.m.	07/2018
Acxiom	a,f	categories	y	no	n.m.	unspecified	categories	register	n.m.	05/2018
IndexExchange	n.m.	aggregated only	US	no	no	13m	categories	website	n	09/2018
Criteo	a	aggregated	y	no	n.m.	13m	61	email/mail	n	05/2018
OpenX	a,f	unspecified	US	n.m.	n.m	unspecified	categories	email	y	05/2018
DataXU	a,b,c,f	behavioural	y	not in EU	n.m.	13m	categories	email	n	06/2018
Lotame	n.m.	unspecified	US	health related	n.m	13m	categories	website	y	09/2018
FreeWheel	a,b,f	unspecified	Y	n.m.	n.m.	18m	categories	email	n	05/2018
Amobee	a,f	unspecified	US	n.m.	y	13m	categories	website	n.m.	06/2018
comScore	a,b,c,f	unspecified	y	n.m.	n.m.	n.m.	categories	website	n.m.	12/2017
spotX	a,f	listed	n.m	n.m	n.m.	18m	65	website	y	unclear
Sovrn	a,c,f	n.m	y	n.m.	y	n.m.	unspecifc	webform	n.m.	05/2018
Sizmek	a,b,c,f	segments	y	not knowingly	n.m.	13m	unspecified	website	mixed	05/2018
Twitter	a,b,c,f	listed	y	not allowed	n.m.	18m	16	account	n	05/2018
Microsoft	a,b,c,f	unspecified	y	y	n.m,	13m	>9	account	n	10/2018
Media Innovation	a	unspecified	US	n	n.m.	14m	partners	n.m.	n.m.	09/2011
Quantcast	a,f	listed	y	not in EU	n.m.	13m	33	website	n.m.	05/2018