Evaluating the Usefulness of Subject Access Requests

Tobias Urban
Norbert Pohlmann
Institute for Internet Security
Westphalian University of
Applied Sciences
Gelsenkirchen, Germany
urban@internet-sicherheit.de
pohlmann@internet-sicherheit.de

Martin Degeling
Dennis Tatang
Thorsten Holz
Ruhr-University Bochum
Horst Görtz Institute for IT
Security
Bochum, Germany
martin.degeling@rub.de
dennis.tatang@rub.de
thorsten.holz@rub.de

Abstract

Ad personalization has been criticized in the past for its privacy implications and a lack of transparency and control. Over the last years, many companies have implemented ways to increase transparency about the data that they collect. Some did so to respond to subject access requests—a right granted to individuals by the new *European Data Protection Regulation* (GDPR). To learn more about the data collected by tracking services, we evaluate how companies respond to *subject access requests*. More specifically, we exercised our *right to access* with 38 companies that had tracked us online. Based on these insights, we perform a survey with 490 participants to evaluate the most common approaches to disclose data.

We find that newly created transparency tools present a variety of information to users ranging from detailed technical logs to high-level segment information. Our results indicate that users do not (yet) know what to learn from the data and mistrust the accuracy of the information shown to them.

Author Keywords

GDPR; subject access requests; usability; online advertisement; transparency

ACM Classification Keywords

K.4.1 [Public Policy Issues]: Privacy

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

Table 1: Overview of the amount of companies that provided the collected data with the corresponding type of such data for both rounds of inquiries (R1/R2).

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

Table 2: Overview and amounts of obstacles set up by companies in both rounds of inquiries (R1/R2).

Introduction

Advertisements are an essential part of modern online services' business models. Successful ad campaigns are expected to reach an audience that is most likely interested in the advertised product or service. To reach these audiences, ad-tech companies build *behavioural user profiles* which can include data like assumed interests in products, demographic information, or clickstream data of websites the users were tracked on.

In the ongoing trend towards more transparency, an increasing number of companies offer ways to users to access collected personal data (e. g., *TripleLift*'s web portal [3]). These tools offer users insights of data collected by the companies about them (e. g., sites they tracked the user on) or data they inferred from such data (e. g., interest segments).

To evaluate the usability of the transparency tools, we requested access to our personal data from 38 companies and analyze the success of these *subject access requests*. Afterwards, we conducted an online user study ($n = 490$) to better understand user needs when it comes to transparency in the online advertising ecosystem.

Subject Access Requests

Prior to our experiment, we identified the 25 most embedded third parties as well as the top 25 third parties that engaged most in *cookie syncing* [2] by visiting the Alexa top 500 list [1]. In total, we identified 36 different companies which we contacted in our experiment.

We contacted companies and tried to exercise our *right to access* and *right to portability*, both granted by the GDPR [4], of the data associated with a cookie ID to evaluate the *subject access request* (SAR) process. 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. The GDPR requires companies to grant users' access to their data within 30 days after their initial request. We considered two deadlines: (1) 30 days and (2) 30 *business* days after the request was made.

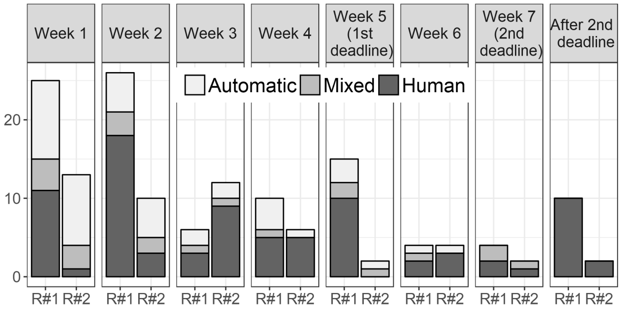
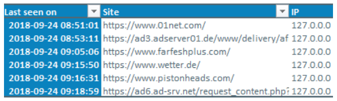


Figure 1: Types and timings of the received responses.

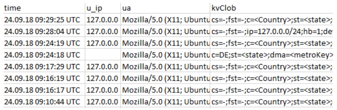
Response Types and Success We grouped responses in three types: (1) *automatic* responses, (2) *mixed* responses, and (3) *human* responses. Figure 1 shows the amounts and types of responses we got during our analysis. In our second round of inquiries, we received fewer responses (approx. half of the amount).

Companies handle inquiries 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. Table 2 gives an overview of the obstacles users face when filling a SAR. Most companies require the user to provide the digital iden-



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

Figure 2: Timestamps, IP addresses, and websites on which the user was tracked.



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>;

Figure 3: Raw (technical) data directly extracted from the user's communication with the company.

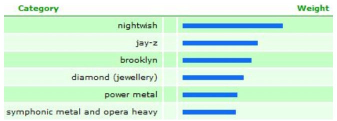


Figure 4: Interest segments inferred from the user's online activities.

tifier (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 interest 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.

It was impossible for the user to know upfront that an affidavit was required since the companies only shared the needed documents via email and did not mention them in their privacy policies. For example, 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, only 21 of 36 companies (54 %) shared data, or told that they do not store any data, 15 (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.

Disclosed Information Table 1 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 extremely heterogeneous in format (e. g., *.pdf*, *.csv*, *.htm*, etc.), data contained (e. g., interest segments assigned to the profile, clickstream data, IP addresses, etc.), and explanation of the data. One company shared an *.csv* file with headers named c_1 to c_{36}, while another company provided detailed explanations in an appended document and yet another told us that we should contact them if we had troubles understanding the data. Some companies shared interest segments they inferred from our (artificially) browsing behavior (e. g., Segment: *Parenting - Millennial Mom*), 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*). Examples for data we received is given in Figures 2, 4, and 3.

After evaluating the SAR process of several companies, we wanted to test whether the provided data is helpful to assess the privacy implications of a company or to take control of the users data. To get an impression of the users' perspective, we conducted an online user survey ($n = 490$).

Results and Analysis

In the survey, participants were shown three data categories sections each listing different data types we received after filing a subject access request. Results of participants views on the provided data are shown in Figure 5. Tools that shared inferred data (i. e., interest segments) were evaluated very positively in all four question categories.

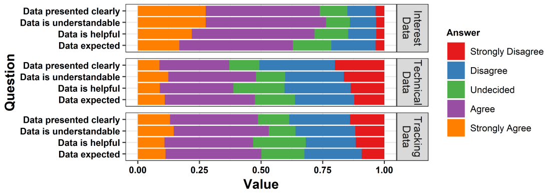


Figure 5: Evaluation of different aspects of the provided data.

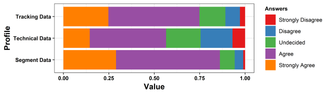


Figure 6: Participants' view on the usefulness of profile categories.

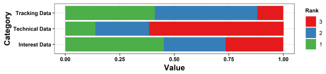


Figure 7: Participants' ranking of profiles

Over three quarters (76 %) of participants stated that they understand ("*Strongly Agree*" and "*Agree*") the provided data and think that it is helpful to understand what companies do with personal data. Fewer Participants (53 %) reported that they understood the tracking data and found it helpful (47 %). Profiles that provide technical data were rated slightly less understandable but much less helpful (39 %), and in these profiles data is presented in a much less clear way (37 %) than in the other profiles categories. We found a correlation between all four questions on clear presentation, understandability, helpfulness, and whether the data was expected, in each section of categories (Pearson-correlation: $r > .5$, $p < .001$). After presenting all three profiles in a section of categories, we asked participants if such data was useful to assess the privacy impact of companies. The results are given in Figure 6. Similar to the assessment of profile categories, segment data was rated to be most helpful followed up by tracking data.

When it comes to preferences which data users would like to receive as response to an access request, participants ranked "tracking" and "segment" data equally as first choice (41 % for tracking and 45 % for segment data) but more users chose "tracking data" as second choice (47 % vs. 28 %—see Figure 7). This was unexpected as participants stated they found segment data to be most useful (see Figure 6) to assess privacy implications of a company and one would expect that they prioritized profiles accordingly.

In general, combined overall profile types, participants who stated that they understood the provided personal data thought that it was useful to assess the use of data ($p < .0001$) and stated it was presented transparently ($p < .0001$). Furthermore, participants who stated that the presented data was useful—regarding usage of data—also stated that the data helped to bring more transparency to the advertisement ecosystem ($p = .0005$).

When asked, 55 % of participants expressed that, after seeing personal data collected on a stranger, that they were *"very interested"* or *"interested"* in data collected about themselves. 58 % stated that they would change their online behavior due to the seen data. Considering that only a few had requested their data previously, this could be related to a social desirability bias, but it could also indicate that there is simply a lack of awareness of transparency tools.

Conclusions Currently, the SAR process of companies is often complicated and tedious. Looking into the response behavior, we see that over 58 % of the companies did not respond within the legal period of 30 days. Participants in our study would struggle to understand and interpret personal data they received if they asked for access, especially if confronted with low-level technical data.

Acknowledgments 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. Alexa.com. 2018. Top Sites for Countries. (2018). https://www.alexa.com/topsites/countries [Online; accessed 2019-05-29].

2. Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. 2018. The Cost of Digital Advertisement. In *The Web Conference 2018 (WWW'18)*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). International World Wide Web Conference Committee, Geneva, Switzerland, 1479–1489.

3. Inc. Triple Lift. 2018. Your Individual Rights. https://access.triplelift.com/. (2018). [Online; accessed 2019-05-29].

4. European Union. 2016. 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 [Online; accessed 2019-05-29].