

**RUHR
UNIVERSITÄT
BOCHUM**

RUB

**Towards Understanding the Impact of the
GDPR on Online Advertisement**
A Technical and Human-Centric Point of View

Tobias Urban

Towards Understanding the Impact of the GDPR on Online Advertisement

A Technical and Human-Centric Point of View

Dissertation zur Erlangung des Grades eines
Doktor-Ingenieurs der Fakultät für Elektrotechnik und
Informationstechnik an der Ruhr-Universität Bochum



vorgelegt von

Tobias Urban

geboren in Bottrop

September 10, 2020

Tag der Mündlichen Prüfung: 03.07.2020

Erstgutachter: Prof. Dr. Thorsten Holz
(Ruhr-Universität Bochum)
Zweitgutachterin: Dr. Nataliia Bielova
(Inria)
Drittgutachter: Prof. Dr. Norbert Pohlmann
(Westfälische Hochschule)

Weitere Kommissionsmitglieder:

Prof. Dr.-Ing. Nils Pohl (Vorsitz), Ruhr-Universität Bochum

Prof. Dr.-Ing. Dorothea Kolossa, Ruhr-Universität Bochum

Prof. Dr. Markus Dürmuth, Ruhr-Universität Bochum

*In loving memory of my dear father Robert.
He would have loved this.*

Abstract

Nowadays, we use the web browser to perform various tasks, e.g., for interacting with our friends, reading news, or sharing our ideas with others. Due to this wide range of applications, web browsers almost substitute operations systems and handle an increasing amount of personal and sensitive data (e.g., credit card numbers or health information). Many of the used web applications relay on online advertisements as one of the primary sources of income because most users want to use services “for free”. Modern advertisements commonly relay on user-specific profiles used to provide them targeted ads. To build these profiles, ad tech companies track users across the Web to understand their habits and personal affectations. Many consider user-tacking to be a privacy-invasive practice as it often happens without direct knowledge or consent of users. The lack of regulation and intransparent data collection and usage patterns resulted in an imbalance of power between service providers (data processors) and users (data subjects) that increased during the last couple of years. To tackle these challenges and to protect the privacy of European Internet users, the *European General Data Protection Regulation* (GDPR) introduced significant changes that affect how personal data can be collected and shared. After a transition period of two years, the GDPR went effective on May 25,

2018. Any company that offers services in the European Union must be compliant with the GDPR—regardless of their headquarters’ location.

This thesis evaluates the effects of the GDPR using a technical and human-centric approach. Generally speaking, we assess challenges service providers face when they want to design GDPR-proof web applications, and we test which changes the GDPR brought to the online advertisement ecosystem on a technical level. Furthermore, we evaluate how users can exercise their right to access and analyze the usefulness of the data provided by these requests.

For our *technical* analysis, we perform two large-scale measurement studies. The first study aims to illuminate third party loading dependencies in modern web applications and determines the deterministic of such. Furthermore, we analyze if any of these parties might conflicting with the new legislation. Our data shows that embedding a single third party might lead to the inclusion of several other third parties. Furthermore, our findings indicate that the embeddings of services on a page load are not always deterministic and that 93 % of the analyzed websites included third parties located in regions that might not be in line with the current legal framework. An important finding of our study is that previous work that mostly focused on landing pages of websites only measured a lower bound as subsites show a significant increase of privacy-invasive tech-

niques. For example, our results show an increase in used cookies by about 36 % when crawling websites more deeply. In the second measurement, we provide a detailed analysis of the information-sharing networks between online advertising companies in terms of cookie syncing. Utilizing graph analysis, we show that the number of sharing 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 leads to a more centralized infrastructure that might have adverse effects on user privacy, as fewer companies perform tracking on more sites.

The *human-centric* analysis consists of two complementary studies. In the first study, we exercise our right to access under GDPR with several companies active in the online advertisement ecosystem and analyze the processes we encountered regarding success, timing, and workload from the user's point of view. We then use the data we received by these requests in our second study to conduct an online survey with 490 participants to evaluate three approaches to disclose data. To complement this view, we shed light on the design decisions and complexities when

building transparency tools in online advertising using an online survey ($n = 24$) and in-person interviews ($n = 8$) with experts from the industry. We observe stark differences between the handing of requests, and disclosure of data: 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. We find that newly created transparency tools present a variety of information to users, from detailed technical logs to high-level interest segment information. Our results indicate that users do not (yet) know what to learn from the data. Furthermore, they mistrust the accuracy of the information shown to them. At the same time, new transparency requirements pose several challenges to an industry that excessively shares data that even they sometimes cannot relate to an individual.

Kurzfassung

Heutzutage benutzen wir das Internet um vielfältige Aufgaben zu erledigen. Zum Beispiel interagieren wir mit unseren Freunden, lesen Nachrichten und teilen Ideen mit anderen. Aufgrund dieser weitreichenden Anwendungsmöglichkeiten haben Web-Browser klassische Betriebssysteme beinahe substituiert. Browser verarbeiten immer mehr persönliche und sensible Daten (z.B. Kreditkartennummern oder Gesundheitsdaten). Viele der genutzten Web-Anwendungen nutzen Online-Werbung als primäre Einkommensquelle, da Nutzer*innen diese meist „umsonst“ nutzen wollen. Moderne Online-Werbung nutzt üblicherweise benutzerspezifische Profile, um Nutzern gezielte Werbung auszuliefern. Zum Erstellen dieser Profile versuchen Werbetreibende die Online-Aktivitäten der Nutzer*innen nachzuvollziehen (engl. „tracking“) umso ihre Vorlieben und Gewohnheiten zu lernen. Dieses Nachverfolgen wird von vielen als Eindringen in die Privatsphäre der Nutzer*innen interpretiert, da es oft ohne Einwilligung oder Wissen der Nutzer*innen geschieht. Das Ergebnis ist ein starkes Machtgefälle zwischen Diensteanbietern (Verantwortlichen) und Nutzern (betroffene Personen), das in den letzten Jahren deutlich zugenommen hat. Zur Lösung dieser Probleme und um die Privatsphäre europäischer Internetnutzer zu schützen hat die *Datenschutz-Grundverordnung* (DSGVO)

signifikante Änderungen wann und wie personenbezogenen Daten verarbeitet werden dürfen eingeführt. Nach einer Übergangsphase von zwei Jahren ist die DSGVO am 25 Mai 2018 in Kraft getreten. Alle Firmen, die Europäern Dienste anbieten müssen sich an die Regelungen der DSGVO halten, egal in welchem Land der Hauptsitz der Firma ist.

Diese Dissertation analysiert den Einfluss der DSGVO mittels eines technischen und eines anwenderorientierten Ansatzes. Es werden die Herausforderungen denen Firmen gegenüberstehen wenn sie Dienste entwickeln wollen, die konform zu der DSGVO sind, analysiert. Des Weiteren werden Änderungen welche das neue Gesetz auf das Ökonomiesystem der Online-Werbung hat evaluiert. Außerdem wird geprüft wie Nutzer*innen von dem Recht auf Datenübertragbarkeit Gebrauch machen können und ob diese die Daten, die sie so erhalten, als nützlich empfinden.

Innerhalb der *technischen* Analyse wurden zwei großangelegte Messstudien durchgeführt. Die erste Studie beschäftigt sich mit den Abhängigkeiten von dritten Parteien in modernen Web-Anwendungen und prüft, ob diese deterministisch eingebunden werden. Die Ergebnisse zeigen, dass das Einbinden einer dritten Partei zum sukzessiven Laden vieler weiterer Parteien führen kann. Des Weiteren zeigen die Messungen, dass nicht immer deterministisch entschieden werden kann welche dritten Parteien geladen werden und dass 93% aller analysierten Webseiten min-

destens einmal eine dritte Partei einbinden, die im Konflikt mit geltendem Recht stehen könnte. Ein weiteres wichtiges Ergebnis der Studie ist, dass die Ergebnisse vorheriger Arbeiten, die nur die Startseite einer Webseite analysiert haben, nur als untere Grenze gesehen werden können, da Unterseiten stärker Techniken einsetzen, die in die Privatsphäre der Nutzer*innen eingreifen. Die Messungen haben zum Beispiel gezeigt, dass ungefähr 36 % mehr Cookies gefunden werden wenn Webseiten tiefgreifender analysiert werden. In der zweiten Messstudie wird der Austausch von Informationen zwischen Werbetreibenden („Cookie Syncing“) detailliert analysiert. Die Messungen zeigen, dass der Austausch, nach in Kraft treten der DSGVO, zwischen den Firmen um ungefähr 40 % zurückgegangen ist. Allerdings nahm das Teilen von Daten im Laufe der Zeit wieder leicht zu. Die Messungen hat keinen Rückgang beim „tracking“ von Nutzern oder in der generellen Struktur wie die Firmen miteinander verbunden sind gezeigt. Die Änderungen in dem Ökonomiesystem haben zu einer zentraleren Infrastruktur geführt, die sogar negative Auswirkungen auf die Privatsphäre der Nutzer*innen habe könnte, da weniger Firmen aktiv sind welche dafür mehr „tracken“.

Die *anwenderorientierte* Analyse besteht aus zwei komplementären Studien. In der ersten Studie wird analysiert inwieweit Nutzer*innen von ihrem DSGVO Recht auf Datenübertragbarkeit Gebrauch machen können. Dazu wurde

bei verschiedenen Firmen, die in der Online-Werbung aktiv sind, das Recht ausgeübt und der Prozess bezüglich des Erfolgs, Aufwandes und des zeitlichen Ablaufs analysiert. Die Daten, die durch diesen Prozess erhalten wurden, werden anschließend in der zweiten Studie genutzt, um drei gängige Varianten der bereitgestellten Daten mittels einer Online-Umfrage mit 490 Teilnehmern zu evaluieren. Diese Sichtweise wird durch Expertenmeinungen in Form von Online-Fragebögen ($n = 24$) und Interviews ($n = 8$) komplementiert, um so Designentscheidungen und Herausforderungen bei der Entwicklung von Werkzeugen zur Steigerung der Transparenz zu verstehen. Innerhalb der Studien wurden starke Unterschiede wie Firmen auf Nutzer*innenanfragen reagieren beobachtet. Nur 21 der 38 angefragten Firmen (55 %) haben innerhalb der von der DSGVO festgelegten Zeitspanne Informationen geteilt und nur 13 (34 %) haben Zugriff auf die gesammelten Daten gewährt. Die geteilten Daten enthalten unterschiedliche Informationen von technischen Rohdaten bis hin zu abgeleiteten Interessen der Nutzer*innen. Die Ergebnisse zeigen auch, dass Nutzer*innen noch nicht wissen, wie sie diese Daten nutzen können und sie zweifeln die Vollständigkeit dieser an. Des Weiteren bereiten diese neuen Anforderungen einer Industrie, die stark auf das Sammeln und Teilen von Daten angewiesen ist, große Probleme weil selbst diese die Daten nicht immer zweifelsfrei zuordnen können.

Acknowledgements

First and foremost, I want to thank Prof. Dr. Norbert Pohlmann for his support not only during this thesis but during my course of study. At the Institute for Internet Security—if(is), he provided me a pleasant and productive research environment while always providing me the freedom to explore my ideas. I benefited a lot from his long experience and expertise, not only on a professional level. Furthermore, I would like to thank my supervisor Prof. Dr. Thorsten Holz, who supported me in the last three years. Thorsten toughed me how to conduct impactful research, and I could greatly profit from his strong background and knowledge. Likewise, I want to thank Martin Degling and Dennis Tatang for their support and very cooperative teamwork and insightful discussions.

Moreover, I want to thank my family and friends that supported me emotionally and always motivated me to keep going. Finally, I wholeheartedly want to thank Anika Schramm. Without her help, patience, understanding, and graciousness, all of this would not have been possible.

The research activities that I conducted when writing this thesis were partially supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia (MKW grant 005-1703-0021 “MEwM”).

CONTENTS

- 1 Introduction** **1**
- 1.1 Motivation 1
- 1.2 Thesis Contributions 3
- 1.3 Thesis Outline 12
- 1.4 Academic Publications 15

- 2 Background & Related Work** **19**
- 2.1 The GDPR 20
- 2.2 Advertising Economy 26
- 2.3 GDPR Literature Survey 28

2.3.1	The Right of Access and Right of Data Portability	28
2.3.1.1	The Authentication Problem of SARs	29
2.3.1.2	Success Rates of SARs	31
2.3.1.3	Data Provided through SARs	34
2.3.1.4	Security Concerns of SARs	35
2.3.1.5	Legal Aspects of SARs	37
2.3.2	Measurement Studies	38
2.3.2.1	Changes in the Tracking Ecosystem	39
2.3.2.2	GDPR Compliance	42
2.4	Further Related Work	48
3	Technical Aspects	53
3.1	Privacy Web Measurements	54
3.1.1	Background on Tracking and Cookies	55
3.1.2	Vertical Measurement Approach	56
3.1.2.1	Terminology	57
3.1.2.2	Website Corpus	58
3.1.2.3	Measurement Framework	60
3.1.2.4	Assessing Cookie Usage	67
3.1.3	Results	69
3.1.3.1	General Overview	70
3.1.3.2	Replication and Comparison	79
3.2	Assessing Third Party Dynamics	88

3.2.1	Building Third Party Trees	89
3.2.2	Analysing Third Party Trees	95
3.2.2.1	Cookies Set in 3 rd Party Trees	96
3.2.2.2	Determinism of Third Parties	100
3.2.2.3	Resulting Tree Depth	102
3.2.2.4	Non-GDPR Adequate Parties	104
3.3	Cookie Syncing and GDPR	108
3.3.1	Background on Cookie Syncing	109
3.3.2	Measurement Approach	112
3.3.2.1	Measurement Framework	113
3.3.2.2	Mapping of Third-Party Re-	
	lations	117
3.3.3	Results and Evaluation	120
3.3.3.1	Third-Party Sharing Ecosys-	
	tem	121
3.3.3.2	Connections of Third Parties	133
3.3.3.3	Case Studies	141
3.4	Discussion	144
3.4.1	Limitations	148
3.4.2	Conclusion	149
4	Human Aspects	153
4.1	Analyzing SAR Implementations	154
4.1.1	The Rights to Access & Data Porta-	
	bility	156
4.1.2	Study Design	158

4.1.2.1	Approach	158
4.1.2.2	Analysis Corpus	159
4.1.2.3	Transparency Requirements	162
4.1.2.4	Assessing the SAR Process	163
4.1.3	Results and Evaluation	168
4.1.3.1	Evaluation of Privacy Policies	168
4.1.3.2	Subject Access Requests . .	175
4.2	Transparency Tools	192
4.2.1	Analysis of Transparency Tools . . .	194
4.2.1.1	Criteria Definition	195
4.2.1.2	Results	201
4.2.2	Perception of Transparency Tools . .	207
4.2.2.1	User Study Design	207
4.2.2.2	Results	212
4.2.3	Business Perspective	222
4.2.3.1	Company Study Design . .	223
4.2.3.2	Results	225
4.3	Discussion	234
4.3.1	Limitations and Ethical Considerations	236
4.3.2	Conclusion	239
5	Conclusion	243
5.1	Possible Impact and Discussion	243
5.2	Future Work	245
5.3	Thesis Conclusion	248

Bibliography	251
List of Figures	307
List of Tables	313
A Surveys	317
A.1 User Survey Questionnaire	317
A.2 Company Survey Questionnaire	325

CHAPTER 1

INTRODUCTION

In this chapter, we motivate the research questions addressed in this work (see Section 1.1), describe the contributions of the thesis (see Section 1.2), and outline the structure of it (see Section 1.3).

1.1 Motivation

In 1995, HTTP cookies were added to the HTTP protocol which allow to store textual data in the format of `key=value` pairs on the client. The initial idea was to

store the client’s state to enable the development of new e-commerce services. For example, cookies allow building shopping carts, which was previously not easily possible. However, cookies are also utilized to track users by assigning a unique identifier to the user (e.g., `user_id=abcd-1234`). As early as in 1996, journalists publicly highlighted the risks cookies pose to users’ privacy [94]. Overtime tracking techniques evolved and became more elaborate as services used new methods to identify users (e.g., “device fingerprinting” [49]) or to store identifiers (e.g., HTTP ETags [15]). In 1999 and early 2000, researches began to anticipate the problem of cookie-based user tracking [83, 206]. Followed by these concerns, online trackers started their efforts to meet the demands of users to increase the protection of their privacy. The tracking and ad tech industry imposed self-regulating approaches to govern the usage of tracking techniques that all failed due to the adoption of key players in the ecosystem (e.g., HTTP “Do Not Track” [71]). As those mostly failed, the development of independent anti-tracking tools started, with limited success [125].

As early as 2011, the first strict privacy law that regulated usage and collection of personal data went effective in South Korea, the *Personal Information Protection Act* (PIPA) [105]. Other countries, regions, or states drafted their privacy laws, following these regulative initiatives. Arguably, the European *General Data Protection Regulation*

(GDPR) [177] draw more public, private, and academic attention to the matter of regulating the usage of personal data than other regulations did. One reason could be the high fines of up to 4% of a company's annual revenue, which was at least a reason for many companies to adjust their practices [188]. Many companies ignored self-regulative approaches and, therefore, it is interesting to see whether the European legislation has a meaningful impact on the collection and processing of personal data or if it will deflagrate like other approaches.

1.2 Thesis Contributions

Generally speaking, the contributions of this thesis can be separated into three parts. First, it provides a survey of recent academic work that analyzes the impact of the GDPR from angles similar to the topic of this thesis. Second, the thesis takes a close look at the technical impact of the GDPR regarding data sharing and transfer of data. Furthermore, it discusses technical challenges that service providers face when they want to design web applications that are compliant with the new regulation. Finally, we take a human-centric approach to understand whether the real-world implementations of the new user rights, which

were introduced by the GDPR, provide benefits or if they are too complex to grasp for most users.

Survey In the first part of this thesis, we conduct an extensive survey of published literature that focuses on the effects of the GDPR, which are relevant to the contributions of this thesis. More precisely, we provide a broad overview of how companies implemented the “right of access” and “right to data portability”, which obstacles users face when exercising their right, and how successful inquiries to companies are. Furthermore, we assess related measurement studies that aim to understand the impact of the GDPR on the Web. Our results indicate that not every aspect of the Web, which was expected to be impacted, is affected by the GDPR and that seemingly small differences in the measurement setups strongly influence the results. These changes might lead to different and not always comparable results.

Third Party Dynamics and the GDPR In this thesis, we perform two large-scale measurement studies. In the first study, we perform a measurement study on 10,000 websites on the Web and analyze relations between third parties. We use the notion of *third party trees* (TPT) as a metric for loading dependencies of all third parties embed-

ded into a given website. More specifically, a TPT contains information on all third parties (TP) observed when visiting a given website and accounts for the loading sequence of each TP. Consider the following example: *adidas.com* embeds a script which loads content from *Adobe* (3rd party). The script again loads a script from *Tealium* (4th party), which also loads a script from *Akamai* (5th party). As a result, a TPT captures the hierarchical structures of third parties on a given website and enables us to study the typical characteristics and dynamic nature of the modern Web. Furthermore, we show that embedding a single TP might result in loading a non-deterministic set of additional TPs, which might pose privacy or security risks. Previous work in this area has analyzed implications of the presence of multiple third parties on websites. Recently, Ikram et al. [87] raised awareness for the problem of implicit trust created by decency chains in website embeddings. Earlier work focused on the extent of tracking (e.g., [37, 169, 50]), or on the used mechanisms (e.g., [49, 2, 108]), and again other works on defense mechanisms (e.g., [131, 140]), or the effectiveness of such (e.g., [125, 120, 61]). We highlight challenges service providers face when they want to account for present third parties on their service.

Furthermore, we want to assess in more detail by whom third parties are embedded into websites and study the extent of control service providers have on them. Most

importantly, we show that previous studies did not measure the extent of the analyzed phenomenon extensively enough and only measured a (not necessarily generalizable) lower bound of included TP content. Our results show a significant increase in used cookies (36 %) and tracking techniques (6 %) on subsites of a domain. This increment is because existing studies typically focused only on the landing page (i.e., they scaled horizontally) or considered a small number of subsites. In contrast, we found that subsites show a significant increase of cookie usage and, therefore, measuring landing pages only is not sufficient (i.e., a vertical scaling is necessary).

Data Sharing and the GDPR In the second measurement study, we seek to provide insights into the effects of the GDPR on the information sharing behavior between ad services. Previous studies have described how companies share identifiers using cookie syncing [49, 1]. Nevertheless, 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. Throughout our experiment, we used different browser profiles to visit more than 2.6 million websites (\varnothing 221,000 in each crawl; 8,000

unique domains) over ten months to identify ID syncing between third parties embedded in these websites. We use graph analysis techniques to measure connections between third parties concerning 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 [41] and that at the same time, the amount of tracking [67] and cookie usage [37] decreased. However, others found that the effects of the GDPR are not as significant in terms of directly embedded third parties [169]. Our analysis shows that one can observe changes if more complex coherences are taken into account. 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.

Implementation of Subject Access Requests In this thesis, 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 (e.g., information regarding profiling). Besides from the detailed overview of different approaches on how to implement *subject access requests* (SARs) in practice; our work provides helpful pointers for companies, privacy advocates, and lawmakers how the GDPR and similar regulation could be improved.

Usefulness of Data provided by Subject Access Requests Inspired by the previously described results, we present a study on the extent of new transparency mechanisms and provide insights into how users and companies struggle with new opportunities and regulations. In a sys-

tem as complex as online advertising with multiple actors sharing and building upon tracking data, there are numerous challenges to *effective transparency* [143, 202, 115]. First, those collecting the data must be aware of what and whose data they directly or indirectly collect through third parties. Second, transparency is not an end in itself, so when providing personal data to data subjects, it has to be contextualized and presented in a way that conveys the essential facts but does not overwhelm the user. We study aspects of both challenges by evaluating the current state of transparency tools and the data provided to users when they request access. While we first focus on the views and needs of users, we also try to understand the challenges companies face when providing transparency.

Overview In the following, we provide an overview of the key contributions of this thesis. An overview of them is also given in Figure 1.1.

I) The literature survey makes the following contributions:

- We provide a brought overview of the literature that focused on the right to access and right to data portability. We find that depending on the sector in which companies are active, the

success rates deeply vary and that companies often require users to expose more data when they exercise their rights.

- We survey the published work that performs privacy measurements to analyze the effects of GDPR on the Internet. On the one hand, we found that not all parts are affected by the GDPR (e.g., third party usage) and, on the other hand, that results of different studies vary and sometimes even oppose each other, which is probably attributable to different measurement setups.

II) This thesis makes the following contributions regarding the technical impact of the GDPR:

- We show that only measuring the traffic generated by landing pages of a website or only a few subsites leads to the risk of only capturing a (potentially limited) subset of the loaded third parties. This limitation implies that the obtained results might be biased and not generalizable. For example, our study indicates that subsites use substantial more cookies (over 45 %) than the site's landing pages.

- Using our data, we try to replicate previous work to test if they only measured an incomplete view of their studied phenomenon and show that most privacy-invasive technologies occur more often on subsites.
- 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 ten months, starting right before the GDPR's enforcement date, we show that the number of links between companies reduces by over 40 %.

III) The human-centric approach contributes to the following findings:

- We requested access to our data from 38 companies and analyzed 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 that the provided data is extraordinarily heterogeneous, and users sometimes have to provide

sensitive information (e.g., copies of identity cards) to access their data.

- To gain further insights, we conduct an online user study ($n = 490$) to understand user needs better when it comes to transparency in the online advertising ecosystem. We found that—if not explicitly stated—users often do not know who collects their data. Furthermore, users struggle to understand the data provided to them.
- Finally, we investigate the perspective of online advertising companies in an online survey ($n = 24$) and in-person interviews ($n = 8$). They acknowledge problems with existing approaches, some inherent to an ecosystem that is not fully aware of the data flow within.

1.3 Thesis Outline

The organization of this thesis goes along with the contributions previously described. Chapter 2 provides detailed descriptions of legislation, ecosystems, and techniques used throughout the remainder of this thesis. First, the aspects of the *General Data Protection Regulation* (GDPR) that

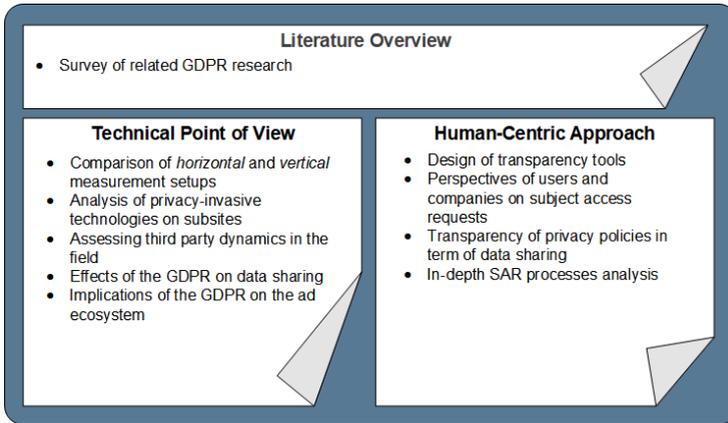


Figure 1.1: Overview of the contributions of this thesis.

are most relevant for this thesis are presented in Section 2.1. A brief explanation of the online advertisement ecosystem is given in Section 2.2. Furthermore, the section presents an extensive survey of related GDPR research (Section 2.3) and an overview of other closely related work (Section 2.4).

Chapter 3 provides an analysis of technical changes, challenges, and effects of the GDPR. The technical part of this thesis starts with a comparison of different measurement setups (Section 3.1) and analyses the effects of different setups regarding the presents of privacy-invasive technologies. Furthermore, Section 3.2 highlights challenges

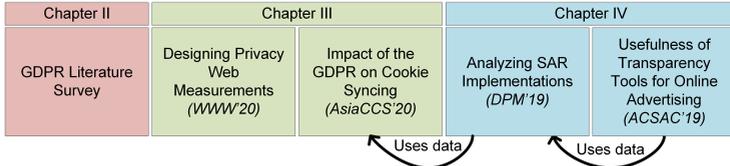


Figure 1.2: Overview of the thesis structure.

service providers face if they want to be compliant with the GDPR in a highly dynamic setting like the Web. In Section 3.3, we analyze the impact of the GDPR on the data-sharing economy.

The second part of this thesis takes a human-centric approach towards the effect of the GDPR (Chapter 4). First, we analyze and compare different approaches companies take when users want to make use of their “right to access” (Section 4.1). Afterward, we perform a user study to test if the data provided by these companies help to estimate the privacy implications of a company (Section 4.2). The thesis closes with the discussion and implications of our findings and incitements for future work (Chapter 5). An overview of the thesis outline is given in Figure 1.2.

1.4 Academic Publications

Several papers, published at competitive and peer-reviewed academic conferences, build the base of this thesis. In the following, we provide an overview of these papers and conferences.

The basis of the topics covered in Chapter 3 are two peer-reviewed papers and a technical report. The first paper was published on the *The Web Conference 2020* (WWW '20), which was held as an online conference [189]. The paper is joint work with Martin Degeling, Thorsten Holz, and Norbert Pohlmann. It discusses conventional web privacy measurement approaches, highlights their limitations, and demonstrates novel results when using different measurement setups. For the paper, I designed the measurement framework and performed major parts of the result analysis. The second paper is joint work with Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann and will be presented at the 15th *ACM ASIA Conference on Computer and Communications Security* (AsiaCCS '20) which will be held in Taipei, Taiwan [191]. The paper discusses the technical impact of the GDPR of data sharing practice in the online advertising ecosystem and measures changes in it. We published additional findings of this work in a technical report [192], which was also the starting point of the second topic discussed in this thesis. In both works, I

designed and performed the measurements, identified the syncing connection, and analyzed—in strong cooperation with Dennis Tatang—the changes in the ecosystem.

Two published papers and, to a lesser extent, on the named technical report build the foundation of the results presented in Chapter 4. An analysis of the technical implementation of new rights granted to users by the GDPR was presented at the 14th *International Workshop on Data Privacy Management* (DPM '19) in Luxembourg and is work in cooperation with Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann [190]. In this work, we make use of our *right to access* and compare different approaches to how companies implement the new right. In this study, I performed the first round of subject access requests and did the complete analysis of the responses. Martin Degeling analyzed the privacy policies. The data provided by the companies lead to direct follow-up research published with the title: “Your Hashed IP Address: Ubuntu—Perspectives on Transparency Tools for Online Advertising”. It was accepted and presented at the 35th *Annual Computer Security Applications Conference* (ACSAC '19) in San Juan, Puerto Rico [188]. This work is again part of the fruitful collaboration with Martin Degeling, Thorsten Holz, and Norbert Pohlmann. The paper analyzes how useful the provided data is to users if they want to assess the privacy impact of different companies. In

the paper, I designed the user study with the guidance of Martin Degeling, conducted all interviews, and performed the analysis of both studies.

Further Publications Aside from the named papers, I published peer-reviewed work not directly related to the topics discussed in this thesis. Most notably is the paper “Towards Understanding Privacy Implications of Adware and Potentially Unwanted Programs,” which won the *Best Paper Award* at the 23rd *European Symposium on Research in Computer Security* (ESORICS ’18) which was held in Barcelona, Spain [193]. The paper provides a first dive into the malicious leakage and collection of private data by different forms of malware. The *Journal of Computer Security* [194] published an extended version of this paper. Both works were done in conjunction with Dennis Tatang, Thorsten Holz, and Norbert Pohlmann. In this work, I designed and implemented the analysis pipeline and performed the analysis together with Dennis Tatang.

CHAPTER 2

BACKGROUND & RELATED WORK

In this chapter, we provide the theoretical background used throughout this thesis and provide an overview of work that is closely related to the research conducted in this thesis. We provide an introduction to the parts of the General Data Protection Regulation that are most important for our work (see Section 2.1) and explain how modern online advertisement works (see Section 2.2). Furthermore, we provide an detailed survey of relevant literature that focused on implications of the GDPR (see Section 2.3). Finally, we give an overview of other related work that was

not conducted with the goal to understand the impact of the GDPR (Section 2.4).

2.1 The General Data Protection Regulation

The *General Data Protection Regulation* (GDPR) [177] is a regulative initiative by the European Union (EU) to harmonize data protection law 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 processing of personal data is allowed, lists rights of data subjects, and obligations for those processing personal data of EU citizens. It applies, therefore, to all companies that offer services that collect and process personal data in Europe. The GDPR was expected to have a substantial impact on the online advertising ecosystem as it provides a broader understanding of what personal data is [154]. Compared to previous legislation, the GDPR also allows data protection authorities to fine companies much higher than before, with up to 4% of their global annual revenue. For example, in one of the first more significant cases, the French data protection authority CNIL (“Commission Nationale de l’Informatique et des Libertés”)

fined Google for €50 million Euros for not validly obtaining consent [35].

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 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. In 2010, the European Data Protection Authorities (Article 29 Working Group) had already decided that profiles created through online tracking are considered personal data and would need explicit consent [38]. However, studies on tracking showed that online advertisers did not follow these recommendations, for example, by ignoring the “Do Not Track” signal [180].

Some tracking companies claim that the data they use is not personal information because it is anonymized, while it is only pseudonymized. If the data were anonymous, it would free them from any data protection related obligations, while attributing pseudonymous data to a person using additional information, still falls under the GDPR’s rules (Recital 26). In 2010, the Article 29 Working Group, a committee of European data protection officials, already made clear 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 [38]. The document concerns the previous directives, but the assessment has been confirmed by court rulings that, e.g., found IP addresses, sometimes also considered pseudonyms, to be personal data. Necessary 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 requests of data subjects (e.g., when they exercise their right to access). According to the GDPR, companies do not have to disclose with which partners they share data but need to mention the categories of partners (e.g., advertisers). However, the French Data Protection Authority CNIL published guidance that companies need to provide a list with which partners they *may* share data [34]. The CNIL has also recently fined Google for not being specific enough in explaining what happens with a user's data [78].

Organizations representing the online advertising industry interpret the legislation differently. For example, the Interactive Advertising Bureau (IAB) Europe has published a working paper for data access requests in April 2018 [64]. They focus on two provisions that limit the data

processor's obligations to answer requests from individuals. First, Article 12 states that a data controller does not need to act on requests if they "*demonstrat[e] that [they are] not in a position to identify the data subject.*" Second, Recital 64 of the GDPR states that data controllers should "*use reasonable measures to verify the identity of a data subject who requests access*" to make sure they do not give away personal information to someone that is just impersonating someone else to get access to information. The argument is that if using the data in a pseudonymous fashion, i.e., a cookie ID, data subjects should have the responsibility to prove that the information connected to that ID is actually about them. Companies could, therefore, decline requests, if the data subject is not the sole user of a browser/device. Also, the guidelines argue that services should make individual decisions as to how they respond to access requests, and should, e.g., not provide clickstream data but information about the segments assigned to an ID.

Article 28 GDPR regulates third-party usage and processing of personal data by third parties. The article states that the service provider ("controller") is responsible for choosing third parties ("processors") that handle personal data according to the GDPR ("*...the controller shall use only processors providing sufficient guarantees to implement appropriate technical and organizational measures [...] and ensure the protection of the rights of the data subject*").

If the third party uses a further partner, the third party is responsible for informing the service provider on the usage of this partner (Article 28 §3 GDPR: “*The processor shall not engage another processor without the prior specific or general written authorization of the controller.*”). However, the service provider is always the responsible contact partner. Therefore, the provider has to handle all user inquiries and might end up having to pay fines when using non-compliant third parties (Article 28 §4 GDPR: “*Where that other processor fails to fulfill its data protection obligations, the initial processor shall remain fully liable to the controller for the performance of that other processor’s obligations.*”). In line with this regulation, an EU Advocate General said that websites that embed the *Facebook* “like button” are jointly responsible for the processing of the data [80]. Hence, service providers are jointly liable to adhere to the new data protection regulation with the embedded third party. However, in the Advocate General opinion, the service providers are only liable for “*those operations for which it effectively co-decides on the means and purposes of the processing of the personal data*”. Hence, service providers should only be accountable for data processing that they can influence (e.g., not for processing performed after the service providers lose control). Advocate General opinions’ are not legally binding but very influential and followed in the majority of cases. Similarly

to this case, the European Court of Justice ruled that providers of “Facebook Fanpages” are joint data controllers with *Facebook* [53].

Besides other rights, the GDPR lists the *right to access* (Article 15 GDPR) and the *right to data portability* (Article 20 GDPR). The difference between those two is that Article 15 GDPR grants users the right to request access to the personal data a company collected about them, while Article 20 GDPR 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 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 answer them within one month (Article 12, No. 3). Companies can extend this time frame by two months. If the data controller needs more time to handle the request, they must state that within a month and give an explanation. Any company that infers information like interests about an individual for advertising purposes performs profiling and needs to disclose this. However, they are not necessarily bound to the additional requirements for profiling mentioned in the respective Article 22, e.g., to enable

human intervention, as profiling for advertising purposes most likely does not have any legal or significant effects.

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 [88] in revenue in the US and €41.8 billion Euros in the EU [85]. The ecosystem behind this is complex and is, in a nutshell, composed out of three entities, which we describe in the following [214, 51]. 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 before videos). 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 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 interests or age.

Free ad space is sold on *real-time bidding* (RTB) platforms whenever a user visits a website. Different entities are involved in the RTB process, but the general flow of information, as described by Yuan et al. [213], is as follows: When a user visits a website, the site provides the available ad space (formally called inventory or impressions) to an ad exchange service, which starts auctions for the available impressions on the site. Websites often use supply-side platforms to provide the inventory. Now, several demand-side platforms place bids on the ad space depending on their estimated value of the impression. They place bids on behalf of the advertisers (e.g., brands) who want to place ads. The highest bid wins the impression, which ensures a maximized ad selling price.

Therefore, user tracking and profiling are critical parts of website and mobile application business models [171, 1, 49]. 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 target advertising [27]. A unique digital identifier is assigned to each user, either by a server or computed based on properties of the user's device (called *device fingerprinting* [49]). The most common way to store such digital identifiers on a user's device is a *HTTP cookie*.

To improve their reach, ad companies utilize *cookie syncing* (sometimes called ID syncing) [142], 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).

2.3 GDPR Literature Survey

In this section, we highlight findings related to the GDPR “right to access” and “right to data portability” (Section 2.3.1) and discuss measurement studies that aim to understand the impact of the new legislation (Section 2.3.2).

2.3.1 The Right of Access and Right of Data Portability

The GDPR [177] introduced seven rights to users: (1) *Right of access by the data subject*, (2) *Right to rectification*, (3) *Right to erasure* (“right to be forgotten”), (4) *Right to restriction of processing*, (5) *Right to data portability*, (6) *Right to object*, and (7) *Automated individual decision-making, including profiling* (Articles 15–22 GDPR). Most

(technical) research only focused on the *Right to data portability* or *Right of access*, respectively, and the *Right to object* in terms of consent management. It is reasonable to limit the analysis to the *right to access*. From a technical point of view, many uses cases are related to this one, from a companies point of view. For example, if a company can identify all data of a user (right of access), it is likely also able to share a copy (right to data portability), to provide the option to modify the data (right to rectification), and to delete the data (“right to be forgotten”). In this section, we survey the literature that focuses on the right of data access and how companies respond to *subject access requests* (SARs).

2.3.1.1 The Authentication Problem of SARs

Depending on the business model of the data processor, user authentication can be challenging or sometimes even impossible. Especially processors that assign unique pseudonym identifiers to users (e.g., tracking IDs) face the problem that they cannot assure that the person performing a SAR presents the own identifier and not the one of another person. In contrast, if a user registers with a service using a digital identifier he or she controls, the service provider can use it to identify the user. For example, if a user provides a cookie ID, there is no way for a company to authenticate

that it is the user’s ID and that she exclusively uses the device. By definition, a cookie ID counts as personally identifiable information, and users have the right to access data associated with it. However, data processors have no option to identify users unambiguously, a Catch-22 for the companies. Recital 57 GDPR proposes that companies should authenticate users by using the same credentials users use to login to their service. However, not all services provide such a feature, and setting them up might be cost-intensive—especially in the online tracking ecosystem. Article 11 GDPR states that data subjects should provide additional information if companies want to identify them, but Recital 57 states that companies should not store any data “*for the sole purpose of complying with any provision of this Regulation*”.

Ad tech companies use different measures to identify users [24, 211]. Some companies solely rely on the companies cookie identifier, others require users to provide additional data (e.g., email addresses), and again other companies require users to sign affidavits that they are the “owner” of the cookie id. The literature calls this practice the “*visibility paradox*” as users have to expose more private data in order to get access to their data [134]. These results show the different interpretations of the new right.

Different studies analyzed in vigorous field studies the authentication approaches of the SAR process of companies

from a wide range of sectors. All studies come to a similar conclusion that companies often require users to present additional information (38 % in one study [211]), that companies lack explanations how users can file a SAR [210], or that it is very tedious and time-consuming to get access as multiple queries are needed because the data processors actively delayed the process [14]. In this thesis, we present a similar finding that companies often make use of most of the defined time frame in which they have to respond to SARs or even determine the start of this time frame autonomously (see Section 4.1). In 2017, a cross-European study analyzed how companies adopted the right of access [133]. As the GDPR was not active at the time, some companies charged users when they performed a SAR (e.g., £10 in the UK), and the time companies needed to respond to SARs varied across countries.

2.3.1.2 Success Rates of SARs

Different studies aimed to analyze the success rates if users exercise their right to access and right of data portability. While the definition of “successful” differs slightly in all works, most define the access to the collected personal data as the main objective of each request. Some works add different secondary goals like the assessment of if the analyzed process is compliant to the new legislation [24] or

if adversaries could abuse SARs [42]. Often the authors of a paper [24, 211, 210] or sometimes educated individuals with a background in privacy performed the SARs [139]. However, there is a lack of knowledge if technically inexperienced users or users without any specific knowledge about Web technologies or online advertisement can perform such requests.

The success rates of SARs differ from study to study; they range from 30–80 % [211, 139]. This aberration seems to depend on the sectors in which the inquired companies are active. In this thesis, we present success rates of around 50 %, in online advertisement (see Section 4.1). An independent work observed similar rates in the same sector [24]. Both works encountered similar obstacles when trying to access their data with these companies (e.g., signing affidavits or disclosing additional data). Furthermore, some companies deny access altogether, or their proposed mechanisms do not work. Studies that surveyed other sectors (e.g., Education, Government, or Finance) reported response rates around 70–85 %, significantly higher than the ones in online advertising [28, 139, 42, 211, 210]. On the one hand, this increase might be due to the challenge that ad tech companies often cannot check the identity of the data subject straightforwardly because they only have a pseudonymous ID, while other sectors might hold other data to identify users (e.g., account numbers). In

Section 4.2, we discuss that, on the other hand, data collection and dataflows in the online advertisement ecosystem are much more complex, which makes it harder for companies to collect all data related to a data subject, and some companies might not be aware of all dataflows. Also, the scale of each study does not seem to impact the results significantly. Studies that performed more than 100 SARs [139, 211, 210] observed similar success rates as studies that analyzed only 14 companies [28]. The time until companies replay to the SARs varies strongly depending on the inquired company. However, often companies either replay within a few days or use the entire 30-day time span (the period defined by the GDPR) [139, 14], which we also observed (see Section 4.1).

Due to the complex process and often unsuccessful outcome of these requests several services emerged that support users performing SARs (e.g., *Dilecy* [45], *TapMyData* [144], or *Datastreams* [40]). Furthermore, tools are being developed that help users to exercise their rights of portability between different services (e.g., the *Data Transfer Project* [39], *My Data Done Right* [23], or *OpenDSR* [138]). These tools aim to make the SAR process simpler for both data subjects and data controllers. However, given the results of the surveyed papers, it is questionable if these tools will improve the status quo, and these tools also come

with the risk that they may get access to sensitive user data.

2.3.1.3 Data Provided through SARs

As previously discussed, SARs are often not successful or might require users to provide additional personal data to get access to their data. However, the most crucial part of each SAR is the data provided as a response as it should bring transparency into the usage of personal data and aid users in assessing the privacy implications of a company. Hence, the provided data should be comprehensive, easily interpretable, and understandable by users. The GDPR specifies that the data should be “*in a structured, commonly used, machine-readable and interoperable format*”. However, it would be favorable if it would also be human-readable.

Different works analyze the data provided through SARs. All works report consistently that the format of the returned data is very heterogeneous in format and included data. Most cases, around 50% in one study, return data is in textual form (e.g., `.csv`, JSON objects, or `.xls`), or in formats that that seems to be intended to be read by humans (e.g., `.htm`, `.pdf`, or email), 30% in the same studies [210, 211]¹. In the remaining cases, the provided dataset

¹Both works seem to be identical but were published on different venues.

seems to depend on the business model of a company or the use case in which it was used (e.g., images or audio recording). In Chapter 4, we provide a detailed analysis of the formats and content of the data returned by ad tech companies. However, accessible ways to display the collected data are not sufficient from a user-centric point of view, and there is a need for novel GUIs too display dataflow, collection, and usage of data [195].

2.3.1.4 Security Concerns of SARs

As subject access requests are an important part for users to get on a level playing field with the data processors, there is an immediate threat that adversaries might abuse SARs [24]. This abuse can either come from the data processors and likewise from users that try to access personal data of other individuals. For example, companies can abuse SARs by forcing individuals to disclose more personal data if they want to access their data or by not providing all the collected data. In Section 2.3.1.1, we already highlighted that several companies collect additional data during the SAR process. Furthermore, previous work reported concerns that some companies do not provide all data upon request [139]. We find similar artifacts (see Section 4.1), and in our user study participants expressed that

they believe that companies withhold some information they collected (see Section 2.3).

It is not necessarily a malicious intention by the companies to not provide all data but that the design of their systems does not all to audit all dataflows. This non-compliance also results from challenges companies face if they want to design systems that enable them to fulfill GDPR requests (e.g., SARs) [82]. Companies need to develop novel systems that aid them in determining all data they should disclose upon request. One approach is to design a platform-independent service that uses a generalized information model to extract the data of interest [84, 26]. However, it seems that small and medium-sized companies have trouble updating their use cases in line with the new legislation, while larger companies are better prepared [164].

If an adversary can fake a user's identity and get access to personal data, she might get access to highly sensitive information (e.g., health records). Requests with malicious intent were proven to be possible using sophisticated methods like faked ID cards [42] or by providing incorrect (fake) data like addresses or birth dates similar to the ones of the data subject [28]. The success rates of reported malicious data access attempts are similar to the ones reported by other works. Hence, companies seem not to check the provided credentials carefully. Besides the negative privacy

implications for users, these malicious SARs also pose a threat to the companies themselves [163]. A motivated adversary might try to use SARs to get information about a companies inner workings (e.g., used technology). For example, the returned data might contain the row names of databases used to store user data, which could potentially help adversaries to plan their attack. *PDGuard* is a technology that aims to tackle several problems that emerge from the misuse of SARs [128]. The framework provides different approaches to mitigate internal and external adversaries that use SARs to get unauthorized access to personal data.

2.3.1.5 Legal Aspects of SARs

Early on, legal scholars raised concerns that the ambiguity of Article 20 GDPR might lead to slow adoption, that it might be costly, and might enable security breaches [199]. Furthermore, the article is not specific when it comes to legitimate collisions of interests between data subject and controllers (e.g., if intellectual property rights might be shared) [73]. Hence, the question arises which data a controller should share upon request. Therefore, it is essential to understand how the legal term “provided” should be interpreted. There is a minimal approach, where only data directly “provided” by the data subject (e.g., data

entered into a form) should be considered and a broader approach, which also labels data observed by the controller (e.g., browser fingerprints) as “provided” [79].

Summary Current research has shown that there are several challenges with the new rights in practice. On the one hand, companies struggle to authenticate users in a way that they do not have to provide additional data or face the problem that adversaries might abuse the system. Clear guidelines and additional research regarding the implementations of systems that resolve these challenges is needed. Such systems might also increase the success rates of SARs and enable more users to exercise their rights. Furthermore, the data provided through SARs does not meet the expectations of users, and there are concerns that it might be incomplete. Overall, the current implementations of the right to access and right to portability seem to be unsatisfactory for both data subjects and data controllers. Table 2.1 provides an overview of the presented papers that analyze technical or legal aspects if users perform SARs or when companies design their SAR processes.

2.3.2 Measurement Studies

In this section, we provide an overview of different studies that aim to measure the technical effects of the GDPR

Table 2.1: Overview of research conducted on subject access requests. The scale of each experiment is given with the respective success rates. An indicator is given if the SAR process (Proc.), the returned data (Data), the user authentication (Auth.), or malicious usage (Mal.) is analyzed.

1 st	Author	Year	Ref	Tech.	Legal	Scale	Success	Proc.	Data	Auth.	Mal.
	This thesis		[190]	✓	✗	38	54 %	✓	✓	✓	✗
	This thesis		[188]	✓	✗	–	–	✗	✓	✗	✗
	Talend	18	[172]	✓	✗	103	30 %	✗	✗	✗	✗
	Wong	18	[210]	✓	✗	230	71 %	✗	✓	✓	✗
	Wong	19	[211]	✓	✗	230	71 %	✗	✓	✓	✗
	Mahieu	18	[139]	✓	✓	106	74 %	✓	✗	✗	✗
	Ausloos	18	[14]	✓	✓	66	44 %	✓	✓	✓	✗
	Martino	19	[42]	✓	✗	55	27 %	✗	✗	✓	✓
	Cagnazzo	19	[28]	✓	✗	19	89 %	✗	✗	✓	✓
	Boniface	19	[24]	✓	✓	30	44 %	✗	✗	✓	✓
	Vanberg	17	[199]	✗	✓	–	–	✓	✓	✗	✓
	De Hert	18	[79]	✗	✓	–	–	✗	✓	✗	✗
	Graef	18	[73]	✗	✓	–	–	✗	✓	✗	✗

(see Section 2.3.2.1) and studies that aim to measure the compliance of services (see Section 2.3.2.2).

2.3.2.1 Changes in the Tracking Ecosystem

While users’ choices regarding online tracking are often not honored [197, 157], it is important to analyze if and how technical aspects of the ecosystem itself changed due to

the new regulatory frameworks. Industry-related groups predicted that the GDPR would cripple online advertisement/tracking [130, 175] while privacy advocates praised the new legislation as a tool that allows users to “regain control of their data”. As these opinions were a positive signal to privacy-aware users, the effect of the GDPR was still unclear.

Before the implementation of the GDPR, only a few companies dominated the tracking ecosystem. These were mainly big players on the Internet (e.g., *Google* or *Facebook*), some CDNs which could potentially track users (e.g., *Amazon* or *Cloudflare*), and some ad networks (e.g., *Adnexus* or *Criteo*) [49, 152, 112]. The GDPR did not change the market-dominating position of the big online tracking companies, but contrary to this, it consolidated their position, especially of *Google* [67, 166, 98]. Furthermore, the new legislation did not impact the connectivity of trackers between each other [166]. Regarding the presence of third parties in general, the changes are marginal and seem not to be directly tied to the GDPR [169, 41]. In this thesis, we provide findings that support these findings (see Section 3.3). However, specific categories of websites show a decline of used third parties (e.g., *News* websites) while others are hardly affected (e.g., websites of private institutions) [169, 184]. The changes of third party usage differ

based on the users' origin even if the same legislation applies in all countries [169, 112, 113].

The ubiquitous presence of third parties is not a direct concern that the GDPR intended to regulate. Instead, the GDPR aims to limit unnoticed collection and usage of personal data. Therefore, it is more interesting to analyze the effects on online tracking specifically. After enforcement of the GDPR, a slight decrease in tracking was observed within the EU while it grew in the US in the same period [67]. In line with the dishonoring of user consent, trackers still track users even after they opted-out of, and some trackers only respect the choice for a short period of time [155].

The studies did not observe a substantial change in the tracking ecosystem itself. However, specific techniques that enable tracking might still be affected by the new legislation. When it comes to the usage of cookies, different studies observed differing changes in the cookie usage of websites. While studies found no change in the usage of third parties [41], another study observed a drop of almost 50% of third party cookies [37]. Other works also observed a decline of cookie usage, to a lesser extent through (around approx. 30%) [98, 12, 112]. Different measurement setups and visited websites are likely a reason for these divergent findings. After the GDPR enforcement, many websites stop using third-party cookies, especially smaller

ones [37]. Since 2002, the ePrivacy directive [176] regulates the usage of tracking techniques. However, most websites largely ignored the directive, even four years after it went effective [184].

2.3.2.2 GDPR Compliance

Aside from effects in the design of applications (e.g., usage of third parties), it is interesting to analyze if services are compliant with the new law. It is worth noticing that there is no definite way to say whether a specific practice (e.g., transferring user data to non-European countries) is compliant to the GDPR as companies can sign *data processing contracts* with their partners ensuring that they handle personal data in line with the legislation. Therefore, the findings of the work that is presented in the following can mostly be seen as indicators or possible problematic dataflows, even if the works sometimes claim differently.

Regarding the transfer of personal data, an essential part of the GDPR is that it may only be transferred within the European Union or to GDPR adequate countries [52]. Most trackers are most of the time compliant with that requirement [92]. Hence, they transferred data mostly within the borders of the European Union. This observation results on the one hand because third parties embedded into a website can successively load further third parties

that might conflict with the GDPR [87] and, on the other hand, attributed to mechanisms like load balancing that, in some cases, might use a server outside the EU. A case study on dataflows from Germany to Russia identified several patterns that lead to the (unnoticed) transfer of user data across borders [147]. All of these patterns include, at some point, an entity located in a country outside the EU (e.g., a JavaScript Library).

Article 8 GDPR, as well as the US *Children’s Online Privacy Protection Act* (COPPA, aim to protect children from online tracking. However, user tracking is present on websites [203, 118] and mobile apps [151] that target children as their primary audience. One problem with this is that children, similar to adults, are not aware of the privacy implications of online tracking [215]. In other areas—that naturally affect very private manners—like health [92, 174] (GDPR adds extra protection to this kind of data), pornography [198], or even political orientation [127] vigorous tracking activities were observed.

Different works aim to understand the unnoticed leakage of personal data in mobile Android applications. A large-scale measurement of over one million apps revealed that a not negligible amount of them might have compliance issues (e.g., by undisclosed usage of location services) [216]. In line with these findings, several apps seem to use personal data in ways not disclosed in their pri-

privacy policies [204]. Previous work, conducted before the GDPR went effective, also observed these inconsistencies between the apps' privacy policies and the actions of these apps [217, 165]. Hence, it seems that the GDPR did not affect the privacy practices of app providers. However, more in-depth analysis is needed to fully understand the effect of the GDPR on the Android ecosystem. Most apps seem to collect location data [216], the device ID, or network information [97]. Apps often use side or covert channels to access personal data, which likely happens without the users' consent or knowledge [150]. The results of these studies help regulators to test for compliance problems and app developers to update their privacy practices or policies accordingly [217, 216].

Different systems were designed to aid and assist companies complying with the GDPR through the measurement of their infrastructure and applications. For a start, the scientific community proposed different frameworks to accomplish this task. For instance, by using temporal logic to audit log files automatically [11, 6]. Furthermore, different methods have been proposed to integrate GDPR compliance checks into the development circle of applications, which developers can audit. This integration is mostly done by extending the current business process modeling frameworks (e.g., UML) [21, 4], by designing new models [9], static code analysis [57], or by defining

formal models against which developers can test their systems [200].

Cookie Consent Banner Degeling et al. analyze different cookie banner notifications and effects of the GDPR on privacy policies [41]. They find that more than half of websites provide a cookie consent notice, but only very few offer users a real choice regarding cookie usage. While different types of banner types exist, the outcome should be similar, namely that specific cookies are not set. In the field, websites often instantly set (tracking) cookies when users visit the website and do not update them once a user made a choice [157]. The technical implementation and enforcement of consent are challenging, but designing informative and useable user interfaces such banners is equally hard and vital. The position of the banner has an impact on user interaction with it, and positioning it at the bottom left of a website/mobile leads to most interaction with the banner [197]. Furthermore, Matte et al. have shown that consent banners sometimes store affirmative consent even if none was given or worst if the user opted-out and nudge users to give consent by using predefined options [117].

Summary Overall, the implementation of the GDPR had no imitate effect on third party usage of websites.

However, specific website categories and some websites hosted in specific countries adjusted their cookie usage. Furthermore, while some studies came to similar conclusions of the effects of the GDPR, the magnitudes of impact vary, and other studies even observe opposite results. Table 2.2 provides an overview of the most essential surveyed literature.

Table 2.2: Overview of literature that aims to measure the effects of the GDPR. The table shows if the study aimed to measure compliance of companies or the impact of the GDPR, the used technology to conduct the measurements, and the respective technologies that were measured.

1 st Author	Year	Ref	Compliance	Impact	Setup	Effect	Tracking	Cookies	3 rd parties	Data
Sørensen	19	[169]	✗	✓	<i>OpenWPM</i>	✗	✗	✗	✓	✗
This thesis		[191]	✗	✓	<i>OpenWPM</i>	✓	✗	✓	✓	✗
Degeling	18	[41]	✓	✓	Proprietary	✓	✓	✓	✗	✗
Libert	18	[112]	✗	✓	<i>webXRay</i>	✓	✗	✓	✓	✗
This thesis		[189]	✗	✗	<i>OpenWPM</i>	✗	✓	✗	✗	✗
Dabrowski	19	[37]	✗	✓	Headless Browser	✓	✓	✗	✗	✗
Rao	19	[147]	✓	—	<i>OpenWPM</i>	—	✗	✗	✗	✗
Iordanou	18	[92]	✓	—	Real Users	—	✓	✗	✗	✗
Qiwei	19	[97]	✓	—	Proprietary	—	✗	✗	✗	✓
Arfelt	19	[11]	✓	—	Proprietary	—	✗	✗	✗	✓
Sanchez-Rola	19	[157]	✓	✓	Manually	✗	✓	✓	✗	✗

2.4 Further Related Work

In this section, we provide an overview of work that is related to findings presented in this thesis, but that does not specifically focus on the GDPR.

Web Measurements An extensive body of work exists that measures online privacy through measurements. In 2016, Englehardt and Narayanan published their work on measuring online tracking [49]. 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 among 13 other types of measurements. They find that many websites use highly sophisticated fingerprinting methods (e.g., based on image rendering) and that most companies participate in cookie syncing. In 2014, Acar et al. [1] conduct another large-scale measurement study. In this paper, the authors examined canvas fingerprinting, evercookies, and the use of cookie syncing. According to their study, 5% of the top 100k websites use canvas fingerprints to identify users. Karaj et al. monitored the online tracking landscape over ten months of real users through a browser extension. They illuminate the online tracking business and argue that more transparency and accountability is needed since users struggle to keep control

of their data [100]. Gonzales et al. presented a large-scale study on the use of HTTP cookies [69]. The authors analyzed more than 5.6 billion HTTP requests over 2.5 months. They show that, in practice, HTTP cookies are much more sophisticated than simple key-value pairs and present an algorithm capable of inferring the format of a cookie with high recall and precision rates. Multiple works [167, 15] analyze a mechanism to respawn a deleted HTTP cookie.

ID Sharing & Ad Networks In addition to the studies referenced in Section 2.3.2, additional work analyzed ad networks and id sharing within them. Falahrastegar et al. investigated the connections between third parties focusing on ID sharing [56]. 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 [20]. They show that 52 advertisers or analytics companies observe over 90% of an average user's online clickstream. The prevalence of cookie syncing is also highlighted by Papadopoulos et al. [141]. The authors conduct a year-long measurement concerning cookie syncing and the costs of delivering ads to users. They found that

97% of users are exposed to cookie syncing, that the top advertisers learn over 25% of all user identifiers, and that ad-related traffic can consume up to 8.5% of users' traffic plans. A method to identify server-side information flow in the ad economy was presented by Bashir et al. [18]. They use re-targeted ads to reveal information flows.

The work of Castelluccia et al. is also related to advertising [30]. The authors introduce a method to filter targeted ads and infer the users' interests from them. Their results indicate that an adversary is capable of reconstructing user profiles even if she has access to a limited amount of ads. Kim et al. recently presented their work on an ad budget attack [102]. They present an attack on targeted advertisers that legally drain the advertiser's ad budget.

Transparency Tools In the following, we provide an overview of related work not previously mentioned in our survey of SAR literature (see Section 2.3.1).

Leon et al. [111] evaluate the usability of several tools to limit online behavioral advertising (OBA). The authors conduct a laboratory study in which participants, for example, use tools to opt-out of OBA and show that all tools at the time had severe usability flaws that lead to the misuse and misunderstanding of such tools. Andreou et al. [8] presented an analysis of ad transparency tools provided

by *Facebook*. In their work, they analyze the messages presented by the social media platform that explains why users see specific ads. They find that these messages are often incomplete and misleading. A web browser extension that gives users a more equitable choice with regards to ad blocking was presented by Parra-Arnau et al. [143]. The extension gives users fine-grained control over the ads they see and helps them to understand how companies use their browsing data. In a study with 40 participants, the authors evaluate the performance of their tool and show that re-targeting is the most common ad strategy. Barford et al. develop a scalable crawler that harvests different ads [16]. They find that ads often rely on user profiles build on clickstream data rather than the visited website. In line with this work, Wills et al. analyze how ad tech companies use personal data to display ads [209] by analyzing the ads shown to users. Melicher et al. investigates the users' perspective on perceived benefits and risks of online tracking by conducting 35 user interviews [124]. They find that users, on the one hand, want more control over tracking but, on the other hand, are unwilling to put effort into actually taking control. Schaub et al. evaluate three different tools used to block online tracking (e.g., *Ghostery*) regarding their effectiveness to inform users that they tracked [158]. One result of their study is that the tools do not manage to inform users about tracking, and some users believe that

the analyzed tools track them. Bashir et al. analyzed “ad preference managers”, a special kind of transparency tool, that allows users to see and edit the segments companies have inferred about them [19]. In a user study, the authors analyze the correctness and compare the composition of such tools. They found that only 27 % of participants state that shown interests are relevant for them.

CHAPTER 3

TECHNICAL ASPECTS OF THE GDPR

In this chapter, we provide an overview of the technical changes in the data-sharing ecosystem (see Section 3.3), the technical challenges companies face when they want to be compliant (see Section 3.2), and discuss how different scaling approaches of Web (privacy) measurements affect specific results (see Section 3.1).

3.1 Designing Privacy Web Measurements

A majority of today's online services are a mix of original content and—to a non-negligible extent—third party resources [169]. Most notably, online advertising is embedded using external resources that display ads to finance these services and to provide them to users free of charge. Third parties include other third parties for various means, e.g., libraries are used to develop services quickly, to decrease loading times, and for analytical purposes. Consequently, this leads to a highly dynamic Web with complicated dependencies among all participants. This trend comes with the drawback that some service providers might not be aware of which third parties are delivered to customers in their name when users interact with their website. Ultimately, third parties can pose risks to users, which is unintended by the service provider. For example, third parties can create security problems (e.g., malvertising [107, 162, 168]), might have negative privacy implications (e.g., trackers [49, 50, 2]), or they can include content that might impact users in other negative ways (e.g., crypto miners [153, 103]). Services themselves reinforce these dynamics as they make use of different sets of third parties in different sections and webpages. For example, news websites often insert

scripts to connect with social media below articles, but not on the actual landing page. This example raises the question of whether previous studies that exclusively measured the landing pages (e.g., [37, 169, 191, 49, 87, 125]) captured a complete and comprehensive view of the analyzed phenomenon. In this section, we perform a measurement study to test the effect of a *vertical* measurement setup on different privacy-invasive technologies, in contrast to a horizontal setup.

3.1.1 Background on Tracking and Cookies

In this section, we provide a brief overview of online tracking methods and HTTP cookies.

Online Tracking

Online user tracking is a widespread phenomenon on the Web [49]. It is used to re-identify users navigating the Web and a crucial part of the modern online advertisement ecosystem as it allows them to provide targeted ads. Techniques to track users can be divided into *stateless* and *stateful* approaches. Typically the tracker assigns an ID to each user and stores it in a cookie on the users' device. An upside of the stateless approaches is not possible to circumvent them by deleting third-party cookies. *Stateless*

approaches use specific attributes of the users' devices to identify them [49, 132, 2, 212, 60, 109] (often called “device fingerprinting”). In contrast, *stateful* approaches use the machine's state to identify users. However, they are more error-prone as device-specific attributes tend to change over time [201, 68].

Cookies

An (HTTP) cookie is textual data, limited in size, which can be locally stored on a client by a server. In theory, cookies contain simple `key=value` pairs used for various means, like storing user preferences or maintained the login status. However, previous work has shown that cookie structures can be way more sophisticated in practice and often contain multiple attributes [69]. The original purpose of cookies is to preserve the client's state over different HTTP sessions (e.g., items in a shopping cart). However, using cookies to store unique user identifiers allows a third party to recognize users between different website visits.

3.1.2 Vertical Measurement Approach

In the following, we describe the setup we use to test the effect of different measurement approaches and to replicate the results of other studies. Furthermore, we use this

data to study the dynamics of the Web on application level (i.e., the browser) to gain insights into the usage of third parties and to illuminate reasons how websites embed third parties (see Section 3.2.1). Our study consists of a multi-stage process in which we (1) build a corpus of websites to visit, (2) use *OpenWPM* [49] to crawl these websites and gather first-party links on these websites, and finally (3) visit the crawled links and log all HTTP traffic, cookie usage, the embedded iframes, and JavaScript calls of interest.

3.1.2.1 Terminology

Before describing our approach, we define the two terms we use throughout this chapter. By *TLD+1* we mean the last part of the hostname following the last dot in it. For example, the URL *https://tools.ietf.org* has *TLD=org*, *hostname=tools.ietf*, and *TLD+1=ietf*. In most cases, *TLD+1* is a “second-level domain”. However, some domain name registries use a second-level hierarchy. For example, New Zealand uses various second level domains for different purposes: *.co.nz* for organisations or *.school.nz* for schools. We identified the TLDs using Python’s *tlsextract* [146] package, which accurately splits generic or country code top-level domains (ccTLD). Furthermore, we distinguish between *landing pages* and *subsites*. A website

is a *subsite* (SB) of a *landing page* (LP) if both share the same TLD+1 but have distinct URLs. Hence, first-party links on landing pages, the page that is usually visited first, lead to subsites. We chose to use the term SB rather than “webpage” to highlight the hierarchical relation between SBs and LPs explicitly.

3.1.2.2 Website Corpus

In our analysis, we use the top 1M *Tranco* list [110], which is an aggregation of four other domain top lists. We used the list generated on 03/26/2019 (ID: W9L9)¹. First, we removed all websites with the same TLD+1 and only kept the one with the highest rank. We did so because we wanted to remove URLs of services that offer users the (almost) same functionality. For example if the list contains *google.com* (rank 1) and *google.co.uk* (rank 4) we would drop *google.co.uk* because both domains share the same TLD+1 (*google*). In total, we removed 607 websites in this step. From the remaining domains, we used the top 10,000 domains and grouped them by the category of their content and also sort them into four different buckets based on their ranking.

We used the *McAfee SmartFilter Internet Database* service to retrieve a list of content categories for the web-

¹Available at <https://tranco-list.eu/list/W9L9>.

sites [121]. We cluster the websites by categories because we want to check if the category of a website has an impact on the usage of cookies and other privacy-invasive technologies. Previous work has shown that, for example, *News* websites utilize more third parties (e.g., ad services) than other categories [169]. In total, the websites in our dataset belong to 85 different categories. An overview of the 15 most prominent categories is given in Figure 3.1. In the remainder, we limit the analyzed categories to the top eight categories and combine all remaining categories in “Other”. Additionally, we group the websites by the following buckets based on the website’s rank in the used list: (1) $1 \leq \text{rank} \leq 100$, (2) $100 < \text{rank} \leq 1,000$, (3) $1,000 < \text{rank} \leq 10,000$, and (4) $10,000 < \text{rank} \leq 100,000$. Due to the removal of duplicate domains, bucket (4) holds these 607 domains, 6.1 % of all visited domains. We use the buckets to test whether the popularity of websites has an impact on the usage of specific technologies.

If not stated otherwise, we use the *one-way analysis of variance* (one-way ANOVA) statistical model to find differences between the analyzed groups (‘rank’ and ‘category’ of a website). In all tests, we use a 95 % confidence interval (i.e., $\alpha = 0.95$).

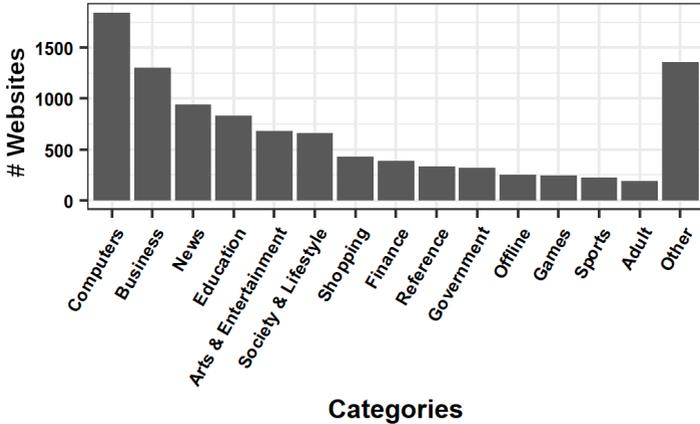


Figure 3.1: Overview of prevalent website categories in our dataset.

3.1.2.3 Measurement Framework

To measure the dynamic of websites, we utilize a customized version of the open-source platform *OpenWPM* [49]. For each visit, we use the same desktop resolution (1366x768) and user agent (Mozilla/5.0 (X11; Linux x86_64; rv:52.0) Gecko/20100101/52.0), allow all third party cookies, do not set the “Do Not Track” HTTP header or other privacy-preserving techniques (e.g., anti-tracking exten-

sions), and use standard bot mitigation techniques to disguise our crawler (i.e., random scrolling and mouse jiggling). Furthermore, the browser adopts other properties from the operating system (Ubuntu 18.04). Aside from our bot mitigation techniques, we do not interact with the visited websites in any way, limitations of this approach are discussed in Section 3.4.1. While a website might detect our crawler, current mechanisms observed in the wild, as presented by Jonker et al. [99], do not detect it.

We configured *OpenWPM* to store all third-party cookies set or accessed via JavaScript and HTTP headers. To capture these events, we instrumented specific JavaScript functions that access the local storage or HTTP cookies, by adjusting the `.prototype` of the respective functions and applying a wrapper to them that logs each call and access to these functions. In contrast to most other studies (e.g., [104, 142, 117]) this allows us to analyze the usage of cookies and not only the setting of them. Furthermore, we inspect all HTTP headers if a cookie is accessed (`Cookie`) or set (`Set-Cookie`). For our measurement study, we disabled Flash because, on the one hand, *Adobe* will deprecate the technology by 2020 [3] and, on the other hand, we did not find a considerable usage ($< 0.01\%$) of Flash cookies in a pre-study we conducted. We describe the pre-study in the following paragraph. We passively log all DNS responses to test if a third party uses an IP address, which

is associated with a country that does not automatically offer a GDPR adequate privacy protection level. We define all countries that are part of the Privacy Shield [179]² and countries part of the European Economic Area (EEA) [72] to be adequate. We use *MaxMind's* GeoIP database [119] to create this association.

Pre-Study As the Web is highly dynamic, any attempt to measure it is quite challenging. To get a comprehensive view of cookie and third party usage, we conducted a pre-study to get an approximation of which measuring parameters to use (e.g., amount of subsites to visit) while limiting the crawling time and generated traffic to a reasonable amount. In the following, we limit our pre-study to third-party (TP) cookies as prior work extensively analyzed those [69, 37, 50, 62, 157, 41, 106], and we want to test whether they might have missed cookies due to their measurement setup. However, in our primary analysis, we also analyze various tracking mechanisms (see Section 3.1.3.2). To find the optimal amount of subsites to visit, we randomly selected 100 websites (TLD+1) from the top 1,000 websites and visited 25, 50, 75, 100, 250, 500, and 1,000 subsites of these websites. The websites were visited in separate

²Note that the experiment was conducted prior to the ruling in the “Schrems II” case [65].

measurements but using the same TLDs+1. We conducted these measurements using a browser with a profile that already has some cookies present in the local cookie store and once with a vanilla browser (e.g., no cookies set) to see if active cookies influence cookie usage. We filled the local cookie store by randomly visiting 100 websites from the top 1,000 websites and used the resulting cookie store. In a separate measurement, we visited the landing page of the selected websites 1,000 times and recorded the used cookies to test if there is a difference if users visit the landing pages or subsites.

We compared the number of TP cookies set in each measurement of the pre-study and found that subsites of websites typically set significantly more cookies than the respective landing page does. In our measurement, the mean amount of cookies used increased by approximately 20 (41%), when visiting subsites rather than only the landing page. This increase shows that if one wants to perform cookie/third party measurements, one should always include subsites to the measurement setup rather than only measuring landing pages. Furthermore, we measured a mean increase of 12 cookies (27%) per website visit when using a browser that already has cookies in the local cookie storage. When it comes to the change of cookie usage based on the number of visited subsites, we found that the mean amount of accessed/set cookies stabilizes around 50

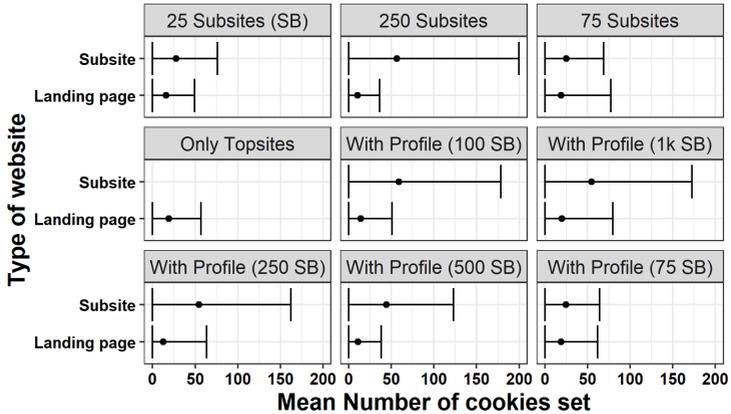


Figure 3.2: Mean number of cookies set in our pre-study with the corresponding standard derivation

(SD: 100; median at 12) after visiting 100 subsites (see Figure 3.2). In conclusion, to magnify the number of cookies set, we use a browser profile that has cookies set and visit 100 subsites and the landing page of each website.

Measurement Sequence We used the same method to create the browser profile for our experiment crawls that we utilized in the pre-study. Before visiting each website, the browser loads this profile but does not alter it. Hence, each

website visit uses the same profile, and the order of visited websites does not impact the results. In total, we conduct the measurements from three different locations (Europe (DE), North America (US), and Asia (JP)) to account for possible geographical differences [37]. For all measurements, we used two computers located at a European university. For each of our regional measurement runs, we created a new distinct browser profile. We used a commercial VPN service (*NordVPN*) to obtain an IP address from the locations outside the EU. Using a VPN service comes with the risk that it might inject content into the communication stream [101]. However, we did not find indications of this practice for the used service. Neither in terms of service nor publicly on the Internet. We conduct our measurements from three areas to get a more comprehensive view on cookie usage and to highlight differences resulting from e.g., different legislation. Furthermore, previous work did find that measuring from just one location might not be enough to get a generalizable perspective [37].

We configured *OpenWPM* to visit the landing page of each website and to gather all first-party hyperlinks on that site (subsites) one day before the first measurement. Therefore, some of these links might not be present on the front page anymore at the time we perform the measurements from different regions or might not exist anymore after all. We did so to increase comparability between our

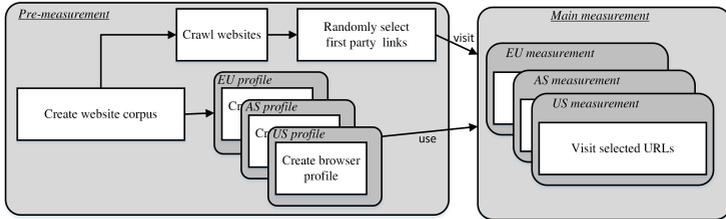


Figure 3.3: Overview of our measurement approach. First, we create the used website corpus, afterwards we create region-specific browser profiles and collect the websites to visit. In the final step, we visit the websites from the different regions and log the traffic.

measurements since we visit the same landing pages and subsites in each measurement. Additionally, we collect all first-party hyperlinks on the subsites but only use them (in random order) if there are not enough subsites linked on the landing page. Afterward, we choose 100 random subsites that we used during the experiment crawls. In each measurement, we visited 549,715 (SD 16,851) distinct URLs on average. An overview of the whole measurement process is given in Figure 3.3.

3.1.2.4 Assessing Cookie Usage

A cookie is a key-value pair set on a client by a visited website or third-party present on that website. Each cookie has a set of attributes, as defined by RFC 2109 [17] (e.g., expiring date or domain associated with the cookie). Meaning, for example, the HTTP cookie header string `uid=abcd; pref_lang=en; expires=Thu, 14 Sep 2020 00:00:00 GMT; Path=/` contains two key-value pairs (i.e., `uid` and `pref_lang`) both with the attributes `expires`, which defines the expiration date of the cookie, and `Path`, which limits the cookie to a specific path on a domain. In this work, we count every single key-value pair as one cookie, and not the entire string, because a TP can use each pair for different purposes. For example, in the previous example to save the preferred language (`pref_lang`) and as a user identifier(`uid`). We extract the keys using Python's `http.cookies.SimpleCookie` library and try to adjust the format of the received cookies, if necessary (e.g., illegal characters or wrongly formatted dates). This process is error-prone as cookies are getting more sophisticated and sometimes do not adhere to the named RFC [69]. Therefore, our results can be seen as a lower bound because we might not be able to extract all key-value pairs. We heuristically group cookies in different categories based on their lifetime. As for HTTP cookies, we compute the lifetime of a cookie

based on the `expires` attribute and the timestamp when the crawler sent/received a request/response, or executed a JavaScript command that set a cookie. If we cannot determine the lifetime of a cookie or if it is negative, we consider a cookie as a “Session” cookie, which is deleted by the browser when the HTTP session ends. By default, items in the local storage do not expire, and thus we classify them as “Permanent” if no expiry date is given. In total, we use four lifetime categories: (1) “Session”, (2) “Short” (≤ 1 week), (3) “Persistent” (≤ 1 year), and (4) “Permanent” (> 1 year). We used the evolution of maximum cookie lifetimes in the Safari browser, enforced through the *Intelligent Tracking Prevention* [10], as an orientation to determine them.

Cookie Classification Cookies can be used for various means. We want to assess the specific purposes why third parties set cookies and which purposes are most dominant to get a better understanding of real-world cookie usage. We use the following cookie type classes defined by the *International Chamber of Commerce UK* [89]:

1. “*Strictly Necessary Cookies*” are needed to provide the basic functionality of a website,
2. “*Performance Cookies*” aggregate (anonymously) the user’s usage of the website,

3. “*Functionality Cookies*” personalize the website’s usage for a user, and
4. “*Targeting/Advertising Cookies*” are used to track users or to display them personalized ads.

For our analysis, we used *Cookiepedia*, a platform that provides public classifications of cookie classes [137]. This process might be error-prone as the service assigns cookie classes by hand, but it is—from our point of view—the best approximation of online cookie usage today. In total, we can classify 45.3% of all observed cookies.

3.1.3 Results

We conducted our measurements in the second quarter of 2019 and found around 93% of the landing pages in our dataset to be accessible. The remaining websites provided services that are not intended to be rendering in a web browser (e.g., APIs) or did not exist anymore. In total, we visited over 1.5 million websites that embedded over 37,000 third parties producing over 4.5 TB of data. More than 17,000 third parties access/set over 59 million cookies across all website visits in our experiment. Table 3.1 gives an overview of our measurements.

Table 3.1: General overview of our three measurement crawls. The number of visited websites and subsites with the corresponding number of observed TPs, cookie setting TPs (C TPs), and used cookies is shown.

Region		Websites	Subsites	TPs	C TPs	Cookies
Europe	(EU)	9,267	561,087	12,076	5,393	20.6M
Asia	(AS)	9,266	530,356	12,926	5,815	18.2M
N. America	(US)	9,333	557,702	13,687	6,115	20.4M

3.1.3.1 General Overview

First, we tested how many subsites set/access cookies in contrast to the respective landing pages to examine the potential bias in previous studies that focused on the landing page only. In our measurements, as shown in Figure 3.4, subsites set considerably more (36%) cookies than the respective landing pages. On average, landing pages set/accessed 55 cookies while subsites set/accessed 78. The difference between the number of cookies used by third parties is statistically significant when comparing (1) different categories (ANOVA test p -value < 0.001) and (2) when comparing landing pages to subsites (p -value < 0.001). However, we did not find a statistically significant effect on the cookie setting behavior by the originating region of the visit or the rank of the website. Our results show that

landing pages of websites show a different cookie usage behavior than the respective subsites as those make more usage of third parties. To get a better understanding of the implications of increased cookie usage, we analyze the primary purposes of cookies.

Lifetime and Cookie Types Aside from the number of cookies set, it is interesting to analyze why websites set them and how long they stay active in the browser as it provides a more detailed picture of the usage and purpose of cookies. Figure 3.5 shows the classifications and computed lifetimes of the observed cookies. Overall, we could classify 45.3% of all observed cookies in terms of distinct used keys. Regarding absolute numbers, we could classify 74% of all observed cookies. Websites use most of the cookies to track visitors or to provide targeted ads (99%). The “type” of the cookie shows a strong correlation with the amount of cookie set for this type (p -value < 0.0001). This result shows that websites set specific types of cookies more often than others. Furthermore, the purpose of a cookie is not related to its lifetime, a X^2 test does not show a correlation between “type” and “lifetime”. Furthermore, third parties use similar types and lifetimes for their cookies, no matter which website embeds them. We did not find a correlation between the “type” or “lifetime” of a

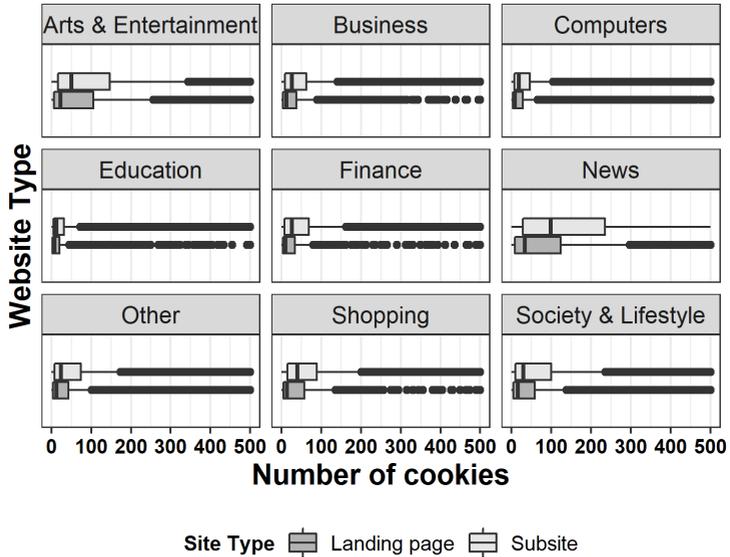


Figure 3.4: Mean number of cookies used by each visited landing page and each respective subsite, by category of the visited website. To increase the readability, we capped the bars at 500. 1.8% of sites had a higher number of cookies; this does not impact the computed values.

cookie and the website’s category. Our results show that user tracking and personalized ads are the overwhelmingly

use cases of cookies. Furthermore, cookies in all categories use various lifetimes. Given the primary purpose of cookies (“Targeting/Advertising”) and the measured increased usage of cookies on subsites, we see that subsites show different behavior in that regard (see also Section 3.1.3.2). Tracking users on subsites provides a more comprehensive view of their online activities. For example, visiting the landing page of an online shop does not necessarily indicate which products a user is interested in, but subsites may provide this information.

Legal Compliance With the introduction of the General Data Protection Regulation and the California Consumer Privacy Act (CCPA) [29], service providers have to be more aware of business partners they work with as they might potentially collect or process personal data of EU citizens visiting the service providers’ websites. If a business partner tracks users or uses personal information in other ways and is not located in a GDPR adequate member state [72] or not a member of the Privacy Shield [179], they need to agree on a data processing contract (Article 28 §3 GDPR) that they take “*appropriate safeguards*” (Article 46 §1 GDPR), which enforce privacy rights of EU citizens. Based on the IP addresses observed in our measurements (see Section 3.1.2.3), we analyzed if third parties estab-

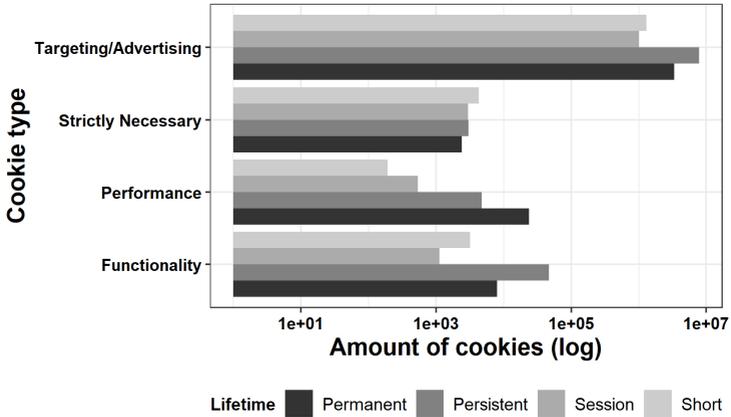


Figure 3.5: Classification of different cookies and the corresponding lifetime. Note the *logarithmic* scale.

lished connections to IP addresses that are associated with countries that are not a member of the EEA or part of the Privacy Shield. In the remainder of the section, we call these parties “non-adequate” or “possibly problematic” to improve the reading flow of this work. Note that every business can agree by contract that the data processing of EU citizens is in line with EU legislation and, therefore, these parties might pose no problem at all (Article 28 §3 GDPR). More precisely, “non-adequate” third parties are parties

that, by definition, do not *automatically* offer a sufficient data protection level. However, the current legal debate only focuses on third parties as “joint controllers” [80, 53] and does not cover fourth or further parties. However, there is no legal certainty if service providers are liable for any actions if they do not know who embedded the problematic object. We want to highlight that a binary classification of what is compliant with legal regulation and what is not is impossible to make without looking at the specific service agreements between websites and third parties.

Figure 3.6 shows the origins and targets of all requests for which service providers need to make sure that they have taken appropriate safeguards. These numbers only refer to our EU measurement, and the results are not violations of the legislation but provide insights into potential dataflows that might conflict with the legal requirements. The origins/targets are based on the observed IP addresses in our measurements. Overall, 4.7% of all cookies were set by services outside adequate geolocations and only 7.1% of the visited domains (TLD+1) exclusively used TPs located at adequate geolocations. Domains using only adequate TPs are located in the US (59%), followed by Germany (7%), and the United Kingdom (3%). In our dataset, Singapore is the most prevalent target of non-adequate requests (26%), followed by China (5%) and Australia

(5%). Percentage-wise, the amount of all requests that target non-adequate geolocations originate in Argentina, Australia (both around 3%), and Austria (1%). The US is the most common origin of such requests (63%), followed by China (6%) and Germany (5%). We did not find a statistically significant impact of the region on the question of whether or not a website uses a third party from non-adequate geolocation. Thus, overall service providers in all observed countries are equally aware of the new legislation. When looking at the services located in possibly non-adequate geolocations, we found that almost half only used sometimes (53%), and the other half always used possibly non-adequate geolocations (47%). Overall, roughly 10% of all observed TPs used IP addresses in possibly problematic geolocations.

In the following, we analyze the services that sometimes use adequate and sometimes non-adequate geolocations. This shifting usage is an exciting subset as service providers might not be aware of the possibility that these TPs change their geolocations over time. In contrast, third parties that always send data to possibly problematic geolocations are easier to identify and, therefore, the transfer of data to these non-adequate countries are likely part of the data processing contracts. Requests to TPs that only sometimes used adequate geolocations were most of the time resolved to an EU IP address but sometimes ($< 1\%$) to addresses

outside the EU. For example, sometimes, a similar resource of a third party was requested from different locations in the same measurement. Meaning, the URL *csm.ad-network.foo* was resolved to *sgp.csm.ad-network.foo* in Singapore and *nl.csm.ad-network.foo* in the Netherlands. These changing geolocations are challenging as service providers cannot ensure that they use EU endpoints of the embedded third party only. In our measurement, *gstatic.com* (a service operated by *Google*) with 20% of all inclusions of possibly non-adequate services and *upravel.com* (a Russian advertising service) with 15% are the top services that might pose a problem to service providers. The next service only accounts for 1% of these possibly conflicting services (i.e., there is a long-tail distribution). One likely explanation is that these are effects of load balancing or similar techniques and that the third party controls all servers associated with the observed IP addresses. However, service providers need to account for this behavior in the data processing contracts with the TP, and the TP must assure that GDPR adequate data processing rules are in place.

Summary Our results show that measuring only landing pages of websites might only reveal a fraction of the websites' real use of third parties. Furthermore, we found that websites make extensive use of cookies, primarily to

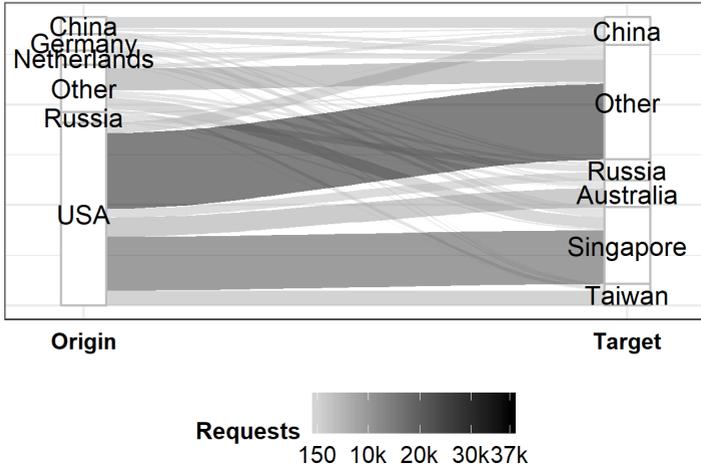


Figure 3.6: Origins (left) and targets (right) of requests to services whose IP address is not mapped to an IP in an adequate country.

serve ads or to track users, and we observed that some embedded TPs might be conflicting with current legislation. To further investigate challenges service providers face and to understand who loads different third parties, we analyze if directly embedded third parties consequentially load further third parties. Additionally, it is also interesting

to look at further areas that might be implicated by our findings.

3.1.3.2 Replication and Comparison

To provide a more comprehensive overview of our measurements in comparison with previous work, we tried to replicate the main findings of previous work using our dataset. We differentiate between studies we could replicate using our data (●—see column “*Rep.*” in Table 3.2) and studies we would partly replicate (◐). Furthermore, we indicate (“*Res.*”) whether we could produce similar results (✓) or not (✗). To reproduce the results, we analyzed the landing pages of each website (if the paper did so) or used the same amount of subsites. If we could replicate the results, we measure them on all visited subsites to test if these studies measured a comprehensive generalizable view or if, as shown in our study, subsites show a different behavior (“*Scales*”). We differentiate if visiting subsites makes a measurable difference in contrast to only visiting landing pages (✗). The results are given in Table 3.2. Our replication studies do not aim to replicate *all* results of previous work, but we only focus on the main takeaways and results closely related to our work. We do not claim that our replications are sound or complete, but we tried to

faithfully replicate previous work as well as possible using our dataset.

In contrast to Dabrowski et al. [37], and as previously stated, we could not find statistical evidence that the originating region of a request influences the cookie setting practices in general. On the one hand, this could be a result of different experimental setups as we tried to maximize the “cookie setting behavior” of each website to achieve more generalizable results. Dabrowski et al. used a headless browser that can be more easily detected by websites and, therefore, might affect the loaded TPs (e.g., ads might not be loaded to counter ad fraud). On the other hand, we performed our experiment on a larger scale and interacted (e.g., scrolling) with the websites, which could fundamentally affect the results.

Table 3.2: Overview of previous work we tried to replicate (Rep.), the scale of the work (“LP” := landing page, “SB” := subsite), the results (Res.) of our replication, and if these experiments show different behavior in a vertical setup (Scales).

1 st Author	Ref.	Year	Venue	Scale	Main finding	Rep.	Res.	Scales
Dabrowski	[37]	2019	PAM	LP	Websites set 49% less cookies if user located in the EU visit them.	●	✗	✓
Sørensen	[169]	2019	WWW	LP + ∅9 SB	Effects of the GDPR to third-party usage is not definite.	●	✓	✗
Sanchez-Rola	[157]	2019	AsiaCCS	LP	Tracking is often still present even if opted-out.	●	✓	✗
Urban	[191]	2020	AsiaCCS	LP + 3–5 SB	Cookie syncing reduced by around 40% after introduction of the GDPR.	●	✓	✓

1 st Author	Ref.	Year	Venue	Scale	Main finding	Rep.	Res.	Scales
Merzdovnik	[125]	2017	EuroS&P	LP + 2 SB	State of the art tracking blocking tools can limit user tracking but still have blind spots.	●	✓	✓
Englehardt	[49]	2016	CCS	LP	Websites use various fingerprinting methods.	○	—	✓
Kumar	[107]	2017	WWW	LP	Implicitly included TPs pose a challenge when upgrading to HTTPS.	●	✓	✗
Ikram	[87]	2019	WWW	LP	Implicitly included parties might pose a security threat.	●	✓	✓
Iordanou	[92]	2018	IMC	user brows- ing be- haviou	In the EU, tracking data is transferred across countries but rarely leaves the EU.	●	✓	✓

1 st Author	Ref.	Year	Venue	Scale	Main finding	Rep.	Res.	Scales
Koop	[104]	2020	PETS	LP + \leq 35		x	—	—
Bashir	[18]	2016	USENIX	LP + \leq 15		x	—	—
Fouad	[61]	2020	PETS	LP + \leq 10		x	—	—

Furthermore, we found that subsites set significantly more cookies than the respective landing pages. As for the results of Sørensen et al. [169], we could verify that the GDPR has no immediate effect on third party usage. Sanchez-Rola et al. [157] show that opting-out of cookies often has no measurable effect on cookie setting practices in the field. We could only partly reproduce this work as we never interacted with any cookie banners, but our results show that cookies are still widely used and that there are no regional differences, while in the EU users should opt-in before a service uses cookies. We used data of our prior work collected before the GDPR became effective (see Section 3.3). Using this data and comparing the regional data in our experiments, we could verify that cookie syncing seems to be influenced by different legislation. Scaled to our collected data, we found an increase in cookie syncing activities on subsites in contrast to landing pages. This replication is not representative as our measurement misses essential features, especially to identify IDs, to assess cookie syncing since we only used one profile in each region.

To test whether our results of increased cookie usage on subsites also applies to user tracking, we use the numbers presented by Merzdovnik et al. [125] on the presence of trackers on websites as a baseline. To test if a tracker is active on a website, we use the *EasyPrivacy* List [47], which is a list combining URLs of known trackers. However, we

do not test whether anti-tracking tools are useful or not. In our measurement, we found that trackers mostly occur on subsites in comparison to their respective landing pages (an increase of approx. 6%). 2.5% of the measured websites do not embed any trackers on the landing page but use trackers on subsites. Overall, we could show that tracking on subsites increases and that future work concerning this area should include subsites into their measurement. In terms of overall tracking occurrence, we produced results comparable to the “plain” profile used by Merzdovnik et al. Finally, we tested the prevalence of device fingerprinting scripts in our dataset, as previously studied by Englehardt et al. [49]. As the scripts identified by Englehardt et al. are probably outdated, we only found four of them in our total dataset. We used the popular “*Fingerprint2*” library [58] to test for the presence of such trackers. Hence, our results are a lower bound as we only test for the presence of one script. We identified a mean increase of device fingerprinting of 25% on subsites in contrast to the respective landing pages. In all three measurements, we found 13 domains (0.14%), which did not use the script on the front page but on its subsites. Overall, we found the tracking script on 0.15% of the landing pages while Englehardt et al. identified device fingerprinting on 1.8%, and the most popular script on 0.45% of the analyzed websites.

The works of Koopet al. [104] and Fouad et al. [61], both published after the defense of this thesis, measured several subsites (up to 35 and 10) to understand the analyzed phenomenon better. As the studies were not available at the time when we conducted this measurement, we did not try to replicate the results (✗). However, given our results and the replication of similar studies (i.e., web-tracking), we expect that measuring more subsites would show a different behavior (i.e., more tracking). In 2016, Bashir et al. performed a measurement study on the dependencies between at tech companies using up to 15 subsites. While the authors made their data available, they did not provide guidance on how to replicate their results on other datasets easily. Furthermore, due to the high dynamic of the Web and especially the ad tech business, the results would not be comparable after all (see the replication of Englehardt’s results). All of the introduced works did not provide reasoning on why they used the specific amount of subsites.

Summary In this section, we demonstrated that only measuring landing pages hides the scale of different phenomena observable on the Web. Furthermore, the behavior of TPs differs on different subsites that raises the question to what extent service providers are in control of TPs em-

bedded into their services. To tackle this challenge, one needs to understand relations between TPs and the determinism of which third parties will be embedded into a service.

3.2 Assessing Third Party Dynamics

Web services make use of resources hosted by third parties for various means. Everyday use cases for third-party usage are libraries used for web development, the integration of social media content (e.g., the *Facebook* “Like Button”), to display ads on websites, or to increase the service’s performance (e.g., using cached fonts). Often these third parties are embedded by adding JavaScript code or an `iframe` element into the website. After injection, these objects perform the desired tasks independently and might even load further resources. For example, an embedded ad might load additional third-party code designed to counter ad fraud, to measure the effectiveness of the ad, or to load additional fonts used by the ad. As a result, embedding a single third party can lead to a long tail of additionally embedded partners.

For this phenomenon, service providers face the challenge to determine the embedded third parties is often only possible on the client at loading time. For example, the dynamic nature of the modern Web is enabled by advertisements that are loaded in real-time to suit the interests of each visitor, content embedded based on the visitor’s location (e.g., cookie banners), or load-balancing mechanisms. Furthermore, new legal regulation pressures services to be more transparent about what data they collect and with

whom they share it. The *General Data Protection Regulation* requires service providers to ensure that personal data from European users are only processed following the law, regardless of the location of the third party. Hence, service providers have a vital interest in knowing which third parties might be embedded into their website, to protect users from potential security and privacy risks, and to comply with current legislation. In this section, we analyze the (non-)deterministic and dynamics of third party inclusions in today's Web.

3.2.1 Building Third Party Trees

In this section, we evaluate the number of partners loaded by an embedded third-party object. To do so, we model *third party trees* (TPTs) for each visited URL (for each landing page and all subsites, respectively), which include all third parties loaded on the visited page. We use the same dataset that we described in Section 3.1 to perform the analysis presented in this section.

We build the trees based on the analysis of elements used to load third-party code dynamically: JavaScript, iframes, and Cascading Style Sheets (CSS). Other HTML objects (e.g., images) can also load third-party content, but these objects cannot load additional code dynamically and would not spawn any children in the tree. In our

analysis, we omit these objects if they are located right below the root ($depth = 0$) but consider them if they occur as leaves in longer branches. We omit them because they would make the results harder to interpret as one cannot decide if these parties do not load further third parties or simply cannot do so. A third party tree shows which party is responsible for loading another party. To account for HTTP redirects, we substitute the respective TLD+1 with the TLD+1 of the redirection target in the trees and delete all edges that would otherwise create a redirection loop. Therefore, we add each loaded script and inserted iframe as a child of the respective ancestor (script/frame) in the tree, if needed. For example, if a script, which is loaded from *foo.com*, loads another script from *bar.com* we add *bar.com* as a child of *foo.com* in the tree. Thus, we can measure the number of third parties loaded due to each embedded object. Regarding iframes, we use *OpenWPM*'s feature to save the nested iframe structure of a website. Based on this structure, we insert each frame (i.e., the source TLD+1) at the corresponding position in the tree. For JavaScript code, we inspect the call stack of each script, test if the script executed code from another party (e.g., a function in an external library), and include this party at the respective position in the TPT (based on the call stack entries). To find CSS dependencies introduced through the `@import` command, we analyze the content type of HTTP

requests and test if the origin and target of the request URL both load CSS. Eventually, each TPT consists of all scripts, style sheets, and iframes loaded by a website. Each branch of a tree represents the embedding sequence of all loaded third parties (domains). Each of these parties could potentially set a cookie and receive personal data of users, at least their IP-address.

If not stated otherwise, we use the TLD+1 of a third party domain as the node identifier; otherwise, we use the companies associated with the TLD+1. We use the *WhoTracks.me* database [33] to link domains to the respective companies owning them. Thus, a branch in the tree could consist of multiple domains operated by the same company (e.g., *foo.com* \rightarrow *googletagmanager.com* \rightarrow *googleapis.com* \rightarrow *youtube.com*). However, we collapsed requests stemming from one company into one leaf. In the previous example, we would not add *googleapis.com* even if *youtube.com* would load a script from that domain. We did so because otherwise, the resulting trees would result in a much deeper length if TP load several resources from the same TLD+1. For example, if *foo.com* was embedded and would then load *metric.foo.com*, subsequently *ad.foo.com* and finally *foo.com/?ad_loaded=1* the resulting branch would be much deeper. Overall, the maximum depth using this more lax approach would increase by magnitudes from eight to 52.

Thus, a branch consists of all TLD+1/companies that could perform a task on the client.

Figure 3.7 provides an example of a third party tree, including the companies' names, not TLD+1s. The tree shows the visited website (*adidas.com*), the directly embedded third parties (*MediaMath*, *TrustArc*, and *Adobe*—*depth* = 0), the partner of the third partners (fourth parties at *depth* = 1—e.g., *Improve Digital*), and further embedded services e.g., *Akamai* (*depth* = 2) or *Instana* (*depth* = 3). We marked the services that actively set cookies with a [C]. The example illustrates that embedding a single service can lead to the loading of many other direct partners into a website (e.g., *MediaMath* embeds four partners). Furthermore, embedding a single third party might implicitly lead to a long branch of direct and indirect partners of the used third party (e.g., *Adobe* that creates a branch of *depth* = 4). Note that at depth four, a service from *Adobe* is embedded. This loading dependency is not a loop, but simply, the previously loaded party utilizes a different service of *Adobe*. All of these parties could set cookies, but not all do.

Distinction from Previous Work In this work, we use the concept of third party trees to analyze the dependencies of TPs and the loading hierarchies among them. A similar concept was used by Ikram et al. [87] and Ku-

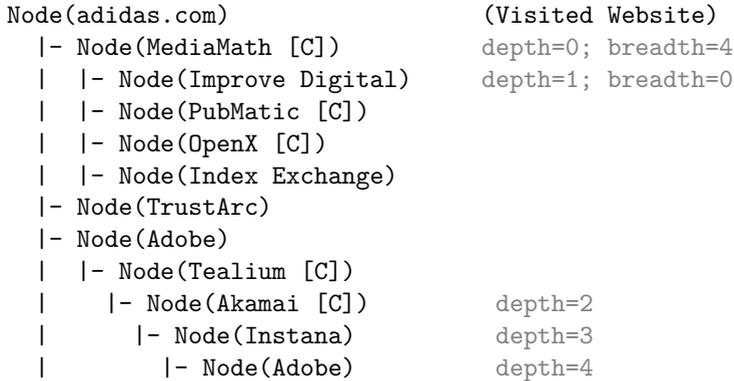


Figure 3.7: Example of an observed third party tree. The listed companies represent the companies operating the observed URLs. [C] illustrates the cookie setting parties.

mar et al. [107] to analyze resource loading dependencies (termed “inclusion chains”). Both works use a concept of the implicit trust of the embedded third and further parties. Kumar et al. show that websites heavily rely on third parties, that almost one-third of websites embed a third party that loads further parties, and that these dependencies are a problem if one wants to serve a website fully via HTTPs. Ikram et al. also show that many websites (approx. 40%) implicitly trust parties loaded by directly embedded third

parties and see an increase in embedded malicious or at least suspicious site or script files in these chains. We extend this concept as we visit several sites of a single domain, which enables us to construct a more comprehensive and realistic view of a website’s dependencies, and we do not limit ourselves to JavaScript inclusions. We use the term *tree* rather than *chain* as our concept describes a complete view of a website’s TP relations and not a single instance of TP inclusion. A similar concept, also called *inclusion chains*, was proposed by Bashir et al. [18] and, in another work, extended to the term *inclusion tree* [20]. Inclusion trees represent the aggregation of loading dependencies of ad tech companies across several websites. While similar in nature, our approach of third-party trees focuses on the service provider’s point of view and not on the perspective of publishers. Hence, we focus on the determinism of third party inclusions on one domain and not the relations between advertisers observable across domains. Furthermore, our work differs from previous work, as most tried to measure effects on a horizontal scale (i.e., visiting a lot of distinct domains) while we instead analyze websites on a vertical scale (i.e., we visit several subsites of the same domain). Furthermore, we focus on privacy-invasive technologies and the determinism of third party dependencies. By this vertical approach and dependency identification, we can (1) analyze if subsites show different behavior com-

pared to landing pages, (2) study effects of embedding different third parties to websites, and (3) understand who is responsible for embedding specific third parties.

3.2.2 Analysing Third Party Trees

As described above, we are interested in understanding dependencies between third parties and possibly resulting in challenges for service providers and users. With our experiments, we aim to understand how websites embed third parties. Therefore, we created *third party trees* (see Section 3.2.1) to better understand the implications of embedding a single third party into a website. Figure 3.8 shows the depth of the measured third party trees by category of the visited websites. Remember that each visited website (i.e., distinct URL) produced its TPT, and the directly embedded third parties are of depth zero. The average third party branch has a depth of one (median also one), and the deepest branch of a tree we found has a depth of eight. The most prevalent depth of all measured branches in all trees is zero (57%). In total, 43.0% of the observed branches have a depth of one or more, which means that these trees include parties that are not necessarily known to the service provider. Therefore, several third parties (in terms of TLDs+1, not distinct companies) load at least one additional partner. Each node in the trees has, on average,

0.9 (SD 37) direct children (breadth) with a maximum of 361, and each branch compromises on average 0.9 (SD 6.4) different companies (max 127). In total, 2,901 TPs (10 %) of the embedded TPs never included any child. In line with the cookie setting practices in general (see Section 3.1.3.1), the originating region has no statistically significant impact on the depth of a TPT. The category of a website has an impact on the depth of a tree (p -value < 0.0001). Similar to the results of previous work, “News” websites tend to use more cookies and third parties [169]. As over 40 % of all TPs load at least one additional partner, it is interesting to look if these use cookies, for example, to track users or to serve them targeted ads.

3.2.2.1 Cookies Set in 3rd Party Trees

Not every party in each TPT, more specifically in each branch, will necessarily set a cookie. Therefore, we analyzed the depth of the cookie setting parties and the overall amount of cookies set in each branch. We limit ourselves to cookies but expect, based on our results presented in Section 3.1.3.2, that other privacy-invasive techniques would likely produce similar results. Starting with the depth of set cookies, on average, 1.5 parties in each branch do *not* set a cookie. In 48 % of all branches no party and only in 125 branches (approx. 0.002 %) all parties set a cookie.

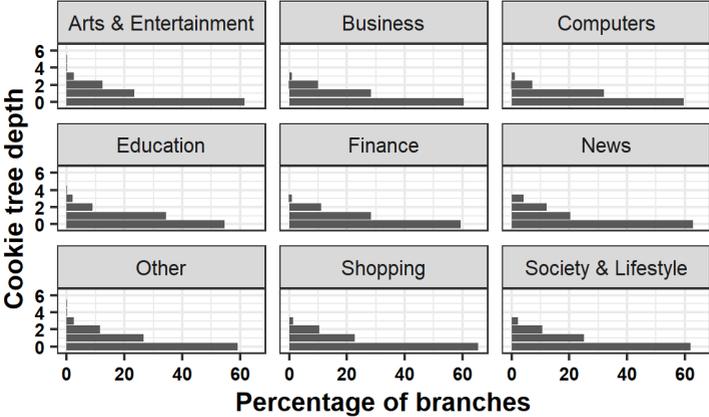


Figure 3.8: Relative distribution of the measured third party tree depth split by the websites' categories.

Again, we found no statistical significance of the region and the number of cookies set in each branch in combination with its depth. The website's category and its rank both show statistical significance in the number of cookies set in each branch (both p -values < 0.001). Figure 3.9 shows the percental amounts of cookies set in each branch. One can see that most companies in the trees (around 68%) do not set a cookie and that the depth of the tree has no significant effect on the perceptual amount of cookie set,

the ANOVA test proofed that to be right. Furthermore, we found that deeper branches do not necessarily, in relative numbers, lead to more cookies usage.

As for the depth of a cookie setting party in the tree, we found that mostly the fourth party ($depth = 1$) sets a cookie (72%). The numbers presented in the figure are scaled for each category, not overall, to account for categories that make more use of cookies than others (e.g., “News”). We found no statistical correlation between the category of a website and the depth of the party that sets/accesses a cookie. The main reason why most used cookies are at a depth of one is likely because most branches are of depth one. Hence, deeper trees occur less often and, consequently, in absolute numbers, set fewer cookies.

Overall, slightly more than 18% of the observed TPs set a cookie at a depth larger one (fifth party or higher). If service providers want to choose services that do not use cookies, for example, because they want to protect their customers from tracking, they face the problem that often the fourth party sets a cookie. Therefore, service providers have to carefully monitor the behavior of all embedded third parties for such behavior. Since one-fifth of the used cookies are at depth more than one, it is worth investigating how much control or knowledge service providers have about these parties. TPs that always include the same third parties can be seen as more predictable because the

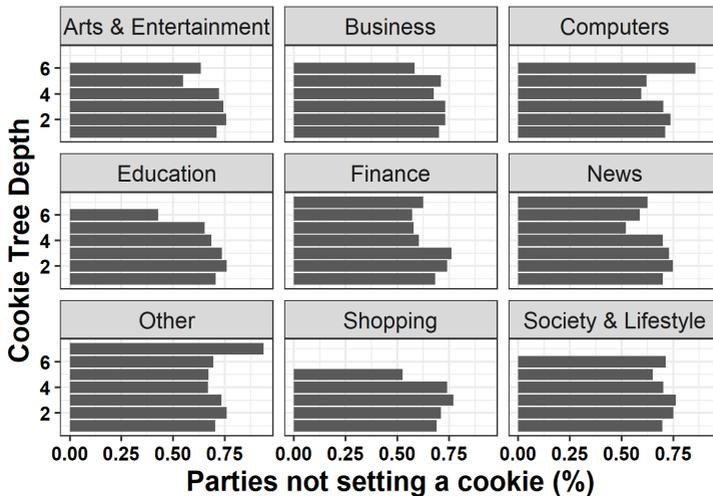


Figure 3.9: Relative amount of cookies *not* set in each branch of the measured trees by category of the visited website (TLD+1).

third parties do not change, and service providers know about the included third parties on their websites. Furthermore, TPs that do not create deep branches are better to assess for service providers since hierarchies and dependencies are easier to understand. Therefore, we analyze the

deterministic of branches generated by directly embedded TPs.

3.2.2.2 Determinism of Third Parties

The determinism of each branch that is generated by an embedded TP is necessary if service providers want to understand which TPs are loaded and who is responsible for loading them. If it is known, before loading the third party object, which other third parties might be embedded, service providers can evaluate the potential risks of a TP for their users. Therefore, we tested the fluctuation of embedded companies for each TP in the measured trees. First, we tested the fluctuation within each visited website (TLD+1) and its subsites. Meaning that we test which third parties are embedded into the visited website by each observed third party on a specific subsite in a specific region. Secondly, we tested the fluctuation across all websites and all regions, meaning that we test if a widespread view of a third party provides more insight of the further loaded parties or if they show different behavior on different websites.

Half of the branches (50.4%) have at least one fluctuating partner in them. Figure 3.10 shows the measured fluctuation of a TP within one visited website (gray) and across all visited sites (black). The x-axis shows the rel-

ative amount of fluctuating companies in all branches of an embedded third party. Zero means branches of this TP always include the same third parties, and six means that six distinct TPs only occurred in some of the branches. These numbers *exclude* third parties that never had any children because these would naturally be zero and might lead to a false conclusion about the deterministic of TPs. The results show that almost two thirds (62%) of third parties that embed other third parties use fluctuating partners (e.g., due to real-time ad bidding) when loaded on different subsites. Across all regions, we see a long tail distribution of companies that only occur in some of the branches, note the increase in more than six new children. Regarding the impact of the originating region, of the performed measurement, we found no statistical significance on its impact on the fluctuation. However, the weighted mean (local) fluctuation was the highest in the US (5.78) and lowest in the EU (5.49).

On a global scale, we find a different picture. We see that the global fluctuation in the EU is more distributed than it is in other regions. We found no statistical evidence that the region affects the local or global fluctuation of children. In conclusion, we see that measuring TPs on a global scale does not necessarily provide a generalizable view as some TPs behave differently on different sites (e.g., due to the advertised products or partners in different

regions). Our results show that the list of third parties embedded in a website is not deterministic, which makes it challenging for service providers to account for all TPs that might be present on their websites. Embedding some third parties leads to an often changing set of embedded third parties (e.g., different TPs providing ads). However, service providers only have little control over these processes as they often depend on third parties to provide their service. As the (non-)deterministic of these trees is related to the embedded TP, it is interesting to analyze the depth of trees generated by different TPs (companies).

3.2.2.3 Resulting Tree Depth

Figure 3.11 shows the average, scaled branch depth that is created by embedding a single object of different companies. All values are scaled for each company, not overall, and include all TLDs+1 operated by the company. Thus, Figure 3.11 presents the resulting depth of each company and does account for the overall occurrence of each company. Furthermore, the figure only lists the top 15 companies, regarding absolute amounts of embeddings of these companies. The category “Other” combines the remaining companies. The top companies account for over 98% of all third-party embeddings. In general, including most TPs results in short trees of depth zero. However, ad-tech

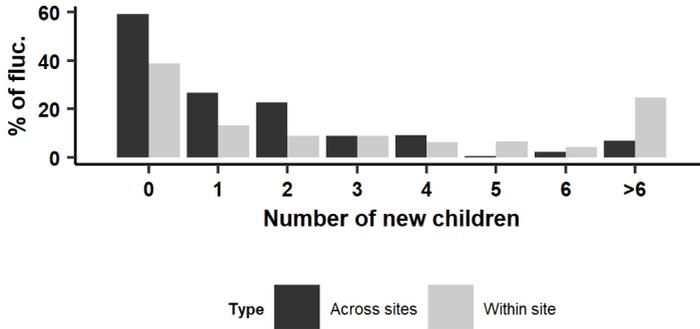


Figure 3.10: Children included in only some of the branches (fluctuation) created by a specific TP within each visited site (grey) and across all sites (black).

companies—the primary source to finance many websites—offer a more widespread resulting TPT depth (e.g., *Pub-Matic* or *Rubicon Project*) which reduces the options to choose partners that do not load many other partners. We found a statistical significance that the embedded company impacts the depth of the generated tree (p -value ≈ 0.008). While the directly embedded TP is responsible for loading further TPs, it is interesting to analyze at which depth of a branch possible conflicting parties occur. Regarding the position of companies in the trees, we found that larger

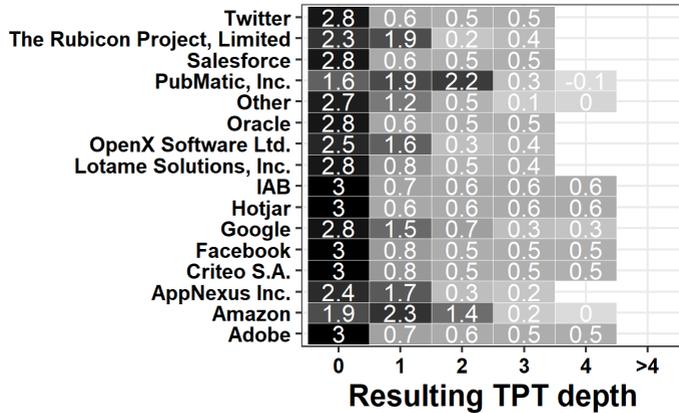


Figure 3.11: Resulting branch depth of objects embedded by different companies (scaled for each individual company).

companies (e.g., *Google* or *Facebook*) occur mostly at depth zero (absolute numbers) while service providers of TPs (e.g., companies that counter ad fraud) occur at deeper levels in the trees.

3.2.2.4 Non-GDPR Adequate Parties

As shown in the previous sections, if a service provider embeds a single third party, they might end up with several

different TPs embedded into their service. Each of these parties could potentially collect and analyze personal data of the service's users. Note that each third party has to inform the first party if they might share data with another partner. The primary purpose of a cookie is to track users or to provide them with targeted ads (see Section 3.1.3.1). In the following, we analyze at which depth in the measured trees cookie non-adequate TPs occur.

Figure 3.12 provides the perceptual amount of non-complaint (black) and the amount of complaint (gray) cookies used in the third party trees. On average, a cookie of a non-adequate party occurs at depth 1.1. Note that the directly embedded third party has depth 0. Thus, these possibly non-adequate parties are, on average, the fourth party. We found that around 10% of the measured non-adequate parties occur on a depth of two or more (1.7% of all cookies). Most cookie setting parties occur at depth one (absolute numbers), but the relative amount of non-adequate cookies is approximately 16% higher than at other depths, excluding depth eight. Depth eight is due to the low amount of occurrences of branches in our measurement not representative (29 in total). We found a significant correlation between the total number of cookies set and the number of possibly GDPR non-adequate cookies set (p -value < 0.0001). Therefore, it seems that the risk of embedding a TP that uses non-compliant cookies is linked

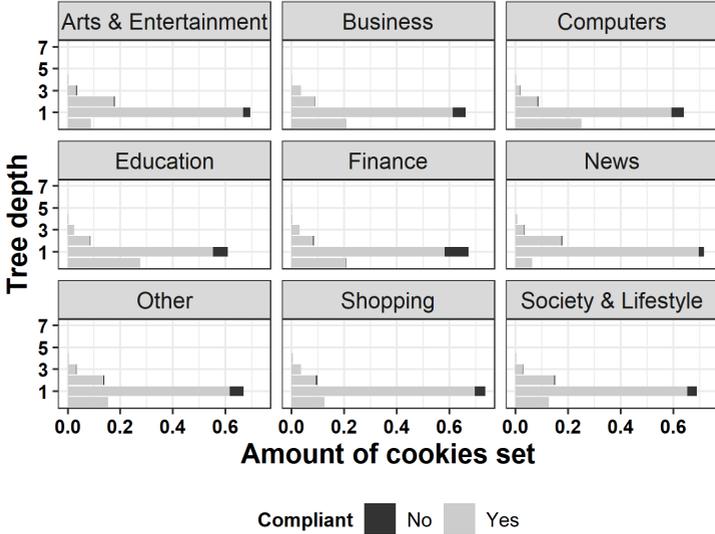


Figure 3.12: Amount of cookies set at each depth of the measured third party trees differentiated by cookies set in GDPR adequate countries (gray) and possibly conflicting countries (black).

to the total amount of embedded third parties (directly or indirectly). Furthermore, the category of the visited website does not show statistical significance regarding the usage of possibly non-adequate parties.

Summary Our results indicate that it is quite challenging for service providers to keep track of all potentially embedded third parties. Furthermore, before loading the directly embedded TP, it is often not definite which other parties might be loaded—especially ad networks load various fluctuating partners. As for parties that are potentially not GDPR adequate, we find that they are most of the time not directly embedded by the service providers but loaded by a directly embedded TP.

3.3 Impact of the GDPR on Cookie Syncing

Advertising remains one of the primary sources of income for many websites, apps, and online services. Many business models rely on ads and analytics services [171] 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 [186]. To provide personalized ads, companies collect data about Internet users through various mechanisms, mainly HTTP cookies [1, 49]. The gathered data is often seen as an economic asset of a company [159]. Nevertheless, attackers also perform malicious exfiltration of personal data [193]. Users are often not aware of the collection, usage, or consequences of the use of their data [36] and have only limited options when trying to control it [157]. To address some of these problems, the *General Data Protection Regulation* introduced significant changes that affect how personal data can be collected and shared. In this section, we seek to provide insights into the effects of the GDPR on the information sharing behavior between ad services. Previous work described how the sharing of identifiers works [49, 1]. However, there is a lack of knowledge about

its extent, the networks behind it, and its development over time.

3.3.1 Background on 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. Storing a unique user identifier in a cookie allows a server to identify a user revisiting a website. It is also common to store additional information exceeding the allowed size for cookies on the server related to that same ID (e.g., inferred-interest segments of users). 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 mechanism prohibits data leakage or cross-domain tracking of different third parties by merely accessing the cookies (via the *Same-Origin Policy*).

If the website originally opened by a user sets a cookie, it is called a *first-party cookie* (A in Figure 3.13). 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 3.13). *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 3.13). 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 (`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 GET 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 of that company but of 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 [136] (see also Section 2.2).

Distinction from Previous Work The introduced related work (see Section 2.3.2) measures the tracking capabilities and other privacy implications of websites—some

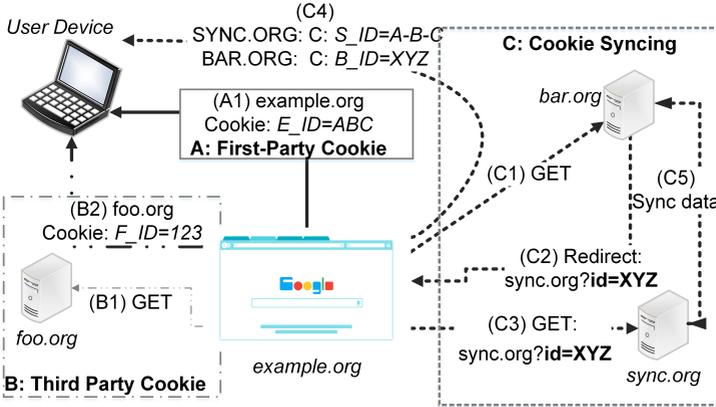


Figure 3.13: Different types of cookies: (A) a first-party cookie—directly set by the visited website, (B) a third-party cookie—set by a third party embedded in the website, and (C) a synchronized cookie—shared between two parties.

concerning the GDPR. However, previous work related to the GDPR looked at third parties present on websites and if their presence changed [169, 67], measured tracking techniques and their prevalence [49, 1], or analyzed cookie setting practices of third parties [41, 157]. In this thesis, we perform a more in-depth analysis and provide insights into the connections of third parties as far as these are observable on the client. We focus on the amount of sharing

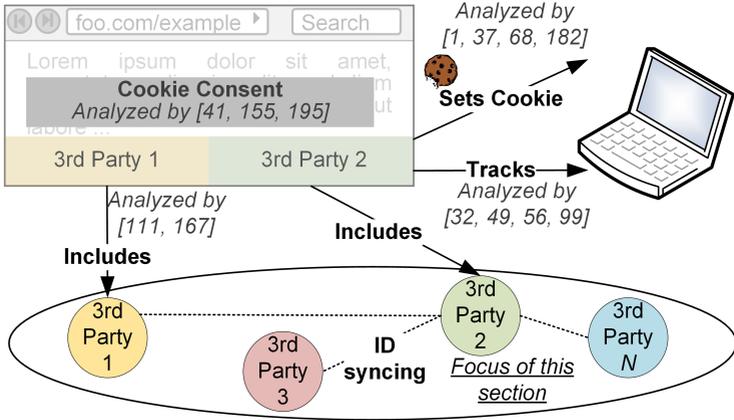


Figure 3.14: Overview of related work and how our work is distinct from it.

connections, the typologies of how companies are arranged among each other, and provide some case studies on specific companies and how they adopted the new legislation. Figure 3.14 highlights our contribution and its distinction from previous work.

3.3.2 Measurement Approach

We conducted a measurement study of cookie syncing in the browser to gain insights into information sharing be-

tween 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.

3.3.2.1 Measurement Framework

To measure the extent of cookie synchronization and the existing networks in the sharing economy, we used the *OpenWPM* [49] platform. For our study, we deployed the platform on two computers at a European university to ensure the 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 crawlers [91]. Additionally, we conducted two measurements using US-based IP addresses using a VPN service to validate the effects of geolocation.

We configured *OpenWPM* 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 allow 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 challenging 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 cookie store, caching, and browsing history. Each profile had its browser storage to make sure that cookies in each session could be stored separately. We created 20 profiles for the top 20 countries with the highest number of Internet users worldwide [90]. 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 top worldwide 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 typical screen resolution³ 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

³We collected the user agents from *TechBlog* [173], most common screen resolution set as reported by *Global Stats* counter [66].

we created by browsing 100 random websites selected from the Alexa top 1,000 list.

For each profile, we took the Alexa top 500 list of the corresponding country [5] (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 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, we visited each website 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. As previously defined, we call these links *subsites* since they are all associate with the same TLD+1 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 subsites. Overall, we visit between 120,000 and 800,000 (\varnothing 221,656; SD 10,609) distinct URLs per

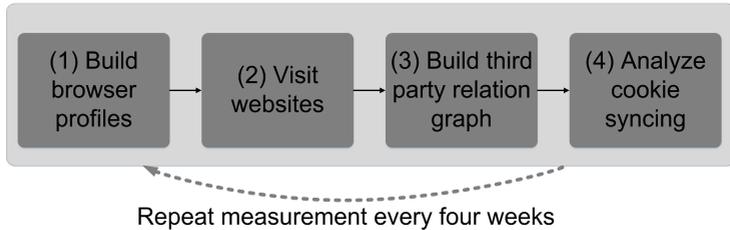


Figure 3.15: Overview of our measurement setup. First, we build the browser profiles which we use to visit the websites. Based on the captured traffic, we build the third-party graphs which we analyze regarding ID syncing.

measurement. An overview of the measurement approach is given in Figure 3.15.

We conducted twelve measurements (M#1–M#12) over 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). In intervals of about four weeks, we made the subsequent measurements (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 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 [101]. In this study, we also used the VPN service (*NordVPN*). To avoid dishonest statements of the VPN service provider regarding the location of their servers [207], we checked at the beginning of each experiment if the VPN service had assigned an IP address associated with a US geolocation using different services (e.g., “*IP Location Finder*” [93] or “*What Is My IP Address*” [208]) 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.

3.3.2.2 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 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 following rules inspired by Acar et al. [1], which are often used to identify IDs (e.g., [49, 61]):

- Eliminate all ID candidates 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., $c1 = (data, 3rw3)$ and $c2 = (data, 70g63b5g)$ would be eliminated).
- Eliminate candidates whose values do not contain enough entropy (according to the Ratcliff/Obershelp pattern recognition algorithm [149]) to contain an ID. Since we only observe a small fraction of the potential ID space, we expect that IDs differ significantly (e.g., we eliminate the candidates $c1 = (id, AAAC)$ and $c2 = (id, AABA)$).
- 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 deflate (e.g., **gzip**) and decode (e.g., **BASE 64**) every HTTP GET and POST argument.

Since any of these arguments might be encoded / inflated multiple times, as previously observed by Starov et al. [171], we repeated this process multiple times, if necessary. We used regular expressions to parse the decoded values for URLs. If we found an URL, we check if this URL has GET parameters that might be an ID, according to our definition of an ID. Other approaches to identify IDs were proposed by Papadopoulos et al. [142] who use a heuristic (by checking if a set cookie is send to party that did not set it) and machine-learning approach (to identify stateless syncing) to identify ID that are synced. Fouad et al. [61] also use proprietary detection mechanism to identify cookie syncing performed by different *Google* services. Both works probably find more syncing attempts than we do and, therefore, our results can be seen as a lower bound.

We used the *WhoTracks.me* database [33] 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 used to sync user IDs, we created a link in the constructed graph for the particular measurement.

3.3.3 Results and Evaluation

To analyze the effects of the GDPR regarding cookie synchronization, 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 sizes of the *OpenWPM* databases. We refer to our first measurement as “pre-GDPR” measurement because we conducted it 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, as we discuss in the following.

Table 3.3 provides an overview of the size of each measurement that varied due to some randomization introduced as described in Section 3.3.2. The table lists the number of domains visited in each measurement to allow comparison of our results with related work. For the remainder of the section, we cluster the observed third parties based on the company operating them. Figure 3.16 illustrates the size of the pre-GDPR measurement concerning the post-GDPR

measurements. While the number of visited domains was above average (of 8,448) but within the interquartile range (25th and 75th percentile), the amount of actually visited websites, in M#1, is above the median (but slightly below the average of 221,656) but also still within the interquartile range.

In line with previous work [41, 169], our data shows that the average number of third parties embedded in websites did not change before and after the GDPR went into effect. Nevertheless, when considering the whole ecosystem, changes can be observed.

3.3.3.1 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 3.3.2 and Table 3.4). All graphs are undirected. Figure 3.17 visualizes graph plots of the first two conducted measurements. Nodes represent companies, and edges represent ID syncing between the companies. Therefore, the nodes reflect the total number of third parties that were embedded in websites and could potentially collect and share personal data. The strength of color and size of node represent sites weight, calculated using the *PageRank* algorithm (the darker and larger, the more critical). Similarly, the color of

Table 3.3: 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	18/05/19	20	8,576	220,948	5.22
M#2	18/05/25	21	8,723	239,636	5.10
M#3	18/06/18	26	8,073	204,108	5.17
M#4	18/07/23	28	8,267	216,283	5.21
M#5	18/08/20	34	8,278	212,405	5.22
M#6	18/09/17	38	8,334	218,687	5.17
M#7	18/10/22	43	8,629	225,230	5.23
M#8	18/11/19	47	8,259	219,164	5.22
M#9	18/12/21	51	8,680	223,718	5.27
M#10	19/01/19	3	8,667	222,122	5.22
M#11	19/02/18	7	8,424	215,407	5.26
M#12	19/03/18	11	8,468	242,165	5.17
∅(2-12)			8,437	221,721	5.20

an edge quantifies its importance (the darker, the more critical). A decrease in the number of nodes means that first parties embed—directly or indirectly—fewer third parties (e.g., trackers or fewer companies participate in the ad bidding process). The amount of edges reflects the number of companies syncing IDs. A smaller number of edges means

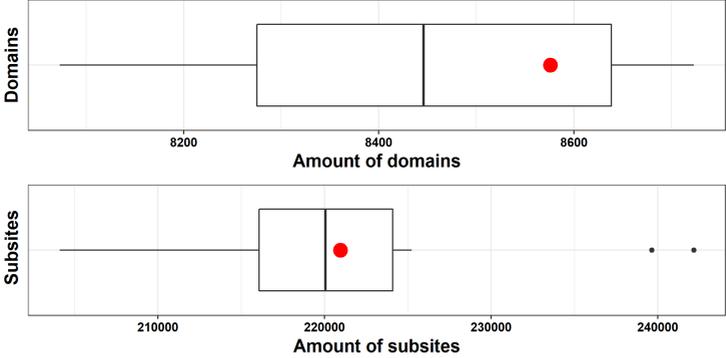


Figure 3.16: Amount of domains and visited subsites used in our measurements (M#1–M#12). The red dots represent the amounts that occurred in M#1.

that fewer companies participate in the sharing economy. The most dominant important node is representing *Google*. Other important nodes represent further companies such as *AppNexus*, *Amazon*, or *Oracle*.

Figures 3.18a and 3.18b 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 light gray dot on the left is the first measurement M#1, in CW 20, before the GDPR came into effect, and the further darker gray (black) dots

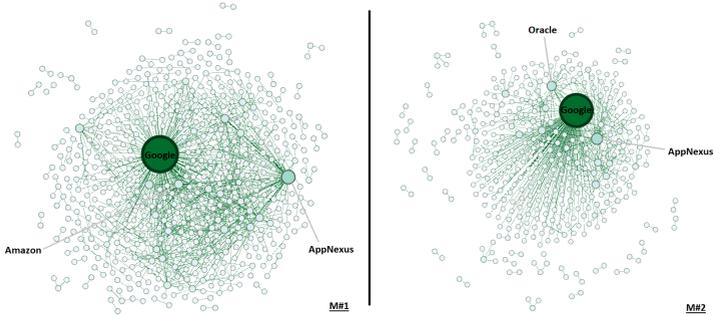


Figure 3.17: The graphs demonstrate the change of syncing connection between our pre-GDPR measurement on May 19, 2018 (M#1, left) and the measurement right after the GDPR went into effect on May 25, 2018 (M#2, right). A reduction of nodes and edges is noticeable—the three most significant nodes are labeled.

represent the corresponding other measurements (M#2 to M#12). We performed two types of linear regression analysis, including the measurement before the GDPR took effect y_{pre} (gray line) and excluding it y_{post} (black 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

Table 3.4: Overview of the measured graph structures (with and w/o isolated nodes) in terms of observed nodes (companies) and connections between them. The relative percentages refer to M#1.

ID	Number of nodes				connections	
	total		w/o iso.		total (w/o 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 %
$\emptyset(2-12)$	10,119	-17.8 %	344	-41.2 %	404	-52.0 %

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 amount of embedded third parties across all website visits between M#2-M#12 are 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 based on the values without the value of the measurement before introducing the GDPR. If the value of our measured pre-GDPR measurement is outside the confidence interval, we confirm that by the time of the introduction of the GDPR, the number of nodes reduced.

The result is 7,151 as the lower confidence limit and 11,774 as the upper confidence limit (see red interval in Figure 3.18a). 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. However, this is a matter of an effect of the GDPR and not directly about the GDPR. The strength of the effect is still minimal, thereby, since the value of M1 lies only barely beyond the interval.

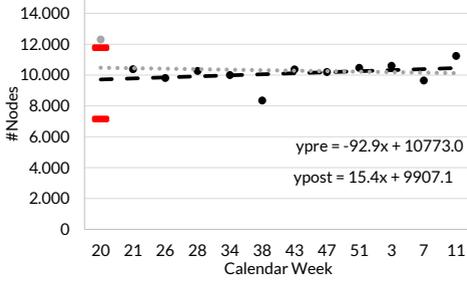
As shown in Table 3.3, the amount of third parties per website stays more or less stable across all measurements, while Figure 3.18a shows a drop of third parties used from M#1 to M#2. However, Table 3.3 lists domain averages and Figure 3.18a shows companies aggregated over all domains. The overall decrease is in line with previous

work that found that websites tended to switch to more prominent ad networks (e.g., *Google* or *Facebook*) when the GDPR took effect [67]. Thus, it is reasonable that the amount of observed independent companies drops (smaller companies disappear), while the amount of used third parties stays stable. We discuss the measured effects on companies active in the ecosystem in Section 3.4.

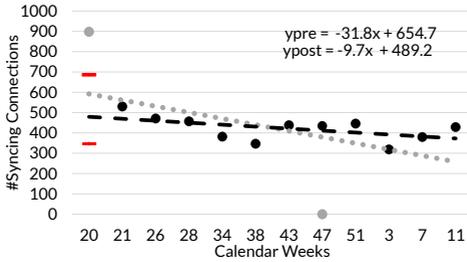
Before the GDPR enforcement, the graph M#1 contained 12,304 nodes, 11,738 of which are isolated. Isolated nodes have no connection to another node and represent third-party companies that are embedded into websites but do not perform cookie syncing (e.g., a JavaScript library). Overall, the number of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

Figure 3.18b shows the numbers of ID sharing connections. Of particular interest is the reduction of syncing relations by about 40% throughout our measurements—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 measurement before the GDPR are both decreasing to different extents.

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 based on the values without the pre-GDPR measurement. The pre-GDPR measurement value (898) is outside the interval (lower limit: 347; upper limit: 686). Thus, we see strong evidence that in M#1 and all other measurements different levels of ID syncing occurred. In this case, the strength of the effect is more evident than with the nodes before.



(a) Number of third parties per measurement



(b) Number of syncing connections per measurement

Figure 3.18: Regression lines of our measurements including the pre-GDPR measurement (gray) and excluding it (black). The red dashes represents the confidence interval (99% confidence) for the prediction for the pre-GDPR measurement point based on all post-GDPR measurements.

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 that 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 3.5 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 paths connect nodes. M#1 has 59 components, with the most significant 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 reduction of the size of the most abundant component by up to 55 % (M#10). However, the median component size remains stable at around two throughout all measurements. The stable size indicates that overall components were not affected by the disappearing connections. However, the number of components dropped.

Similarly, the algebraic connectivity is a measure for the number of nodes and the number of connections between

them. This value can be interpreted as the robustness of the graph concerning the connections. The lower the value, the fewer connections are present. The values of the algebraic connectivity 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%). Due to the internal structure of the ecosystem, we did not measure a significant effect on the structure of our measured graphs.

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 3.18a and 3.18b). Over the month following the introduction of the GDPR, the number of nodes increases again slightly, whereas the number of edges continues to decrease. Therefore, a quadratic function can theoretically represent the number of nodes.

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% above the amount for similar visits from the EU, in a comparable time (CW43 and CW5—which were conducted one week before the US measurements).

Table 3.5: 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 %
$\emptyset(2-12)$	37	−37.4 %	258	−39.9 %	0.0875	−27.2 %

Table 3.6 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` [22] 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. Hence, the underlying ecosystem remains unchanged.

3.3.3.2 Connections of Third Parties

To get a better understanding of the described effects on the overall ecosystem, we analyze the structure of the measured third party graphs. Therefore, we look at the degree of each node and classify them based on the number of *direct* and *indirect* partners. Primary partners are those for which we observed a direct syncing relation, while secondary partners are those with a higher degree of separation. We classified third parties (nodes) in three category types: (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 somewhat balanced amount of primary and secondary partners. We labeled a node

Table 3.6: Characteristics of our graphs w/o isolated nodes.

ID	diameter	median degree	modularity	\varnothing clustering coeff.	\varnothing 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
$\varnothing(2-12)$	9	2.35	0.68	0.10	3.49	46

“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 dataset contains 21 central nodes and 30 balanced nodes. The remaining nodes in the graph are outer corners in a star.

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

We also analyzed the effects of the mean betweenness centrality (b&c) 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. On the contrary, the outer nodes would have a betweenness centrality index of zero because they are only the start/end of the shortest path but never part of the path itself. In contrast to the degree of a node, the betweenness centrality is a factor to measure the links between two topologies. Hence, a high betweenness centrality score shows that a node connects

different syncing communities (i.e., a “bridge”). We computed the betweenness centrality index using the `NetworkX` Python package [75]. With regards to the users’ privacy, a node with higher b&c score has potentially access to data of more companies because it is well connected or at least part of many shortest paths between them. Hence, in the complex sharing economy they will get earlier access to user data and might be able to assign them with higher precision to the correct user (e.g., due to erroneous linking of IDs somewhere along the path).

Similar to our syncing connection regression, we performed a linear regression of the mean betweenness centrality and found a statistically significant decrease in the betweenness centrality ($\alpha = 0.01$ with p -value $< .001$). In extreme cases, the betweenness centrality dropped by up to 60% (mean 30% SD: 11%). Table 3.7 presents an overview of the betweenness centrality properties of our measured graphs. 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. Hence, we can compare if the number of companies that would belong to the top 25% of well connected nodes in M#1 changes.

In line with the findings that the amount of syncing connections decreases, the mean/max betweenness centrality

also decreases. Furthermore, the number of well-connected nodes ($b\&c \geq 5.87$ in our case) and connected nodes decreases, which means that fewer nodes sync with each other—in terms of IDs synced.

The result of fewer companies participating in ID sharing has different effects on the importance of different nodes—in terms of sharing connections and information flowing through the nodes. The most important node’s, *Google*, betweenness centrality 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 did lose betweenness centrality ($\leq 50\%$), and the betweenness centrality of 31 companies decreased significantly by more than 50% (these numbers only include companies 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 that “outer” nodes have low (or zero) betweenness centrality scores. In our scenario, the betweenness centrality is a metric of how prevalent a company is in the syncing ecosystem. Thus,

Table 3.7: Betweenness centrality properties of our measured graphs and the change 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	1,821	33,978	262	15	104
M#3	227	1,507	26,277	251	15	88
M#4	259	1,711	29,197	235	12	100
M#5	191	1,336	22,043	228	13	75
M#6	253	1,153	16,496	186	14	93
M#7	248	1,524	26,523	244	22	95
M#8	274	1,678	28,886	254	8	93
M#9	278	1,715	32,135	285	18	92
M#10	151	862	13,525	214	21	67
M#11	248	1,375	21,714	234	17	78
M#12	271	1,562	25,832	246	14	88
∅(2-12)	240	1,477	25,146	240	15	88

these companies connect to all outer corners of the star as secondary partners. However, we did not see a paradigm change in how companies sync user IDs (e.g., how they are arranged). Overall, the degree distribution in our measured graphs did not vary a lot between all graphs, but the total

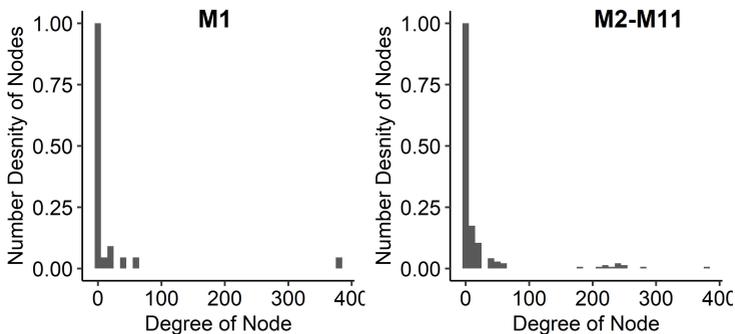


Figure 3.19: Overview of the distribution of the measured degrees of all nodes (excluding isolated notes).

amount of links dropped by 23%. Figure 3.19 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%.

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, but we have shown that ID syncing dropped significantly. Throughout our study, we observed that the number of primary partners of most companies continuously decreased by up to 40% (83 fewer primary partners). Five companies got

isolated, and only two companies gained primary partners. Concerning secondary connections, we see a fluctuation of partners. One explanation for this fluctuation can be 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 a single third party into a website puts users at risk that their data gets shared with hundreds of companies. This exposure leads to the problem that users cannot verify who has received a copy of the data about them and leads to the questions of how service providers can ensure to delete that data upon request. Previous work conducted before the GDPR went effective found that, on average, 3.5 partners get access to an ID (via syncing) [141]. Our measurements have shown that the average amount of ID syncing partners might not be a useful metric (due to the star-like topology) to assess ID syncing, but 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. This finding is in line with our observation of the communities in the graph (see Table 3.5) and public announcements of companies to build tracking infrastructures aside from *Google* or *Facebook* [135].

3.3.3.3 Case Studies

We observed only 70 companies in all 11 measurements. Most of those are prominent companies that offer multiple services (e.g., *Google* or *Oracle*). We found 20 companies (approx. 3%) that shared data before May 28, 2018, and not in any of the consecutive EU measurements, but still appeared in our US measurements. Manual inspection of these services showed that some 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 are required to adopt a technology based on consent management platforms⁴ (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. One other company announced in early 2018 that they are refocusing their business on contextual advertising, where ads are based on the content of a website and not

⁴See <https://advertisingconsent.eu/> for details

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 are sharing with a smaller number of partners, which is in line with other studies that have shown that the reach of smaller companies decreased, while tracking by the market leader increased [32]. An alternative explanation for our results is that companies changed how the exchange IDs. Our measurement approach (see Section 3.3.2) relies on ID syncing observable on the client-side. Therefore, it is still possible that a shift towards server-side ID syncing is taking place, which is not analyzable with current methods. Previous work found that *Google* is one of the beneficiaries of the GDPR, in terms of websites that embed one of their services [32]. Regarding the amount of information flowing through nodes, in terms of ID sharing, we cannot confirm these findings. In 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. Note that our results are not contrary to the findings in previous work but complementary: previous work has shown that *Google* has increased its reach (numbers of websites directly embedding their services),

and our results show that information flow through *Google* (by other third parties) is reduced.

Summary Our results highlight that the GDPR does not affect all entities and parts of the online advertisement ecosystem. For instance, the inclusion of third parties is not as affected as cookie syncing. Furthermore, our results show that the same business practices are still in use and that the general structure of the ecosystem did not fundamentally change. However, larger companies seem to benefit from the GDPR as they increased their reach and market share.

3.4 Discussion

We have shown the challenges service providers face when they rely on third-party code and try to account which third parties are loaded when users visit their service. Furthermore, we highlighted the pitfalls that arise when they comply with current legislation. It is the high dynamic and previously nominal regulation of the Web that now presents challenges to service providers. However, as service providers carefully select the directly embedded third parties (e.g., ad networks), they cannot control which further third parties these included TPs embed once all loaded their content (e.g., due to ad real-time bidding). From a research perspective, we have shown that a simple horizontal scaling of websites to visit (i.e., websites from a given top list) is not sufficient to measure a phenomenon of interest. Meaning that future work should (1) scale their experiments vertically, and (2) previous results of different Web measurement areas should be re-visited to measure the given challenges adequately. Finally, our assessment of purposes why TP uses cookies underlines the dire need of privacy protection mechanisms to limit cookie-based tracking—which is currently promoted by several browser vendors (e.g., Firefox [129] and Safari [10]).

Third Party Dynamics Due to the architecture of the Web and most of the state-of-the-art approaches to build web applications, solving some of these challenges is not easy. One example is that technically service providers can only account which parties are loaded if the client sends them the loaded parties and the party responsible for loading them. Such a mechanism would come with several additional challenges on a technical, performance, and privacy level. On a technical level, service providers would have to implement similar measures, as we did in this study, to account for all third party and to log the TPTs. As we could instrument the browser and JavaScript engine, this task might be harder for the service provider that has to embed such codes in the website itself, with the respective limitations (e.g., JavaScript sandboxing). Performance-wise such measures would increase the traffic on the users' devices as they have to send the data needed for accounting back to the service provider. Finally, service providers get access to personal data of its users as they, for example, might log what kind of ads specific users see, this kind of information might hold sensitive data [30].

Overall, it is questionable if service providers can implement such measures. One possible solution could be to define *Content Security Policies* (CSP) that only allow third parties mentioned in the data processing contracts of the service providers. However, such approaches limit the

dynamic of a website and might negatively impact users using browser extensions [76] or to extensive whitelisting might pose security problems as the CSPs can be bypassed more easily [205]. Another solution might be to forbid fourth party code and to request that third parties directly dynamically embed the further loaded code into their own (i.e., serve as a proxy for all needed third party code).

Cookie Syncing After our first measurement, conducted before the GDPR took effect, we observed a statistically significant drop of ID sharing connections of over 40% within the online advertising ecosystem. The change is likely related to the GDPR that imposed more definite 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 new technology for ID sharing. We did not measure other error-prone syncing attempts in the back end between two companies (e.g., based on IP addresses or device fingerprints [201]). To the best of our knowledge, cookie syncing is still the most common way to share user identifiers. Note that we do not attempt to measure when companies share all of its collected personal data at once with another company (e.g., *Facebook* sharing all of the collected data with other

companies [181]) but instead 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 increased since then. This trend could be an indicator that some websites temporally stop the use of some services but, over time, but 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 [63, 32]. 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 might be 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 essential part of the ecosystem—significantly decreased. However, it did not revolutionize it as the structures remained intact, nor did it dispatch the ecosystem as some industry-related groups pessimistically forecasted [130, 175]. More importantly, the effects on Internet users' privacy might be adverse as fewer companies continue to be present on more websites, increasing their possibilities to create profiles.

3.4.1 Limitations

In the following, we discuss the limitations of the work presented in this chapter.

Third Party Measurement We use the classification of *Cookiopedia*, which might be wrong to some extent and is incomplete. We could only classify slightly over 45% of all observed cookies, but show companies use an overwhelming majority (99%) to track users or serve them targeted advertisements. We mapped requests from different services to a single company, if possible. If we observed multiple requests to domains owned by one company (e.g., `ads.foo.com` and `fonts.foo.com`), we collapsed them to a single request if they occurred in sequence. Our measurement platform, a customized *OpenWPM* instance, does not interact with any cookie banners that are present on the visited websites. Hence, we do not capture cookies set by third parties that honor opt-in choices of (European) users. However, previous works found that cookie banners often do not work as expected [157], do not offer opt-out choices while instead assume opt-in [197] showed that the used consent libraries do not meet other GDPR requirements [41], and that some websites register positive consent even if the user does not express it [117]. Besides, Utz et al. [197] have also shown that the majority of users do not interact with cookie con-

sent notices. Therefore, our results are a lower bound since (1) we shortened the TPTs, and (2) some cookies might only be used after affirmative action of the user.

Cookie Syncing Measurement 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). Again, 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 are a lower bound.

3.4.2 Conclusion

In this chapter, we provided our findings on the cookie setting practices of the top 10k websites in the wild, evaluated third party usage and dynamics, measured the impact of the GDPR on cookie syncing, and discuss how horizontal measurement scaling effects the results of a study.

Measurement Setup We found that 99% of all cookies that we could classify, were set with the intention to track users or to serve them targeted ads. Furthermore, we modeled *third party trees* that assemble all third parties embedded into a website and the loading dependencies among them. By analyzing the third party trees, we found that the median depth of such trees is one (max eight), that there is a sever fluctuation of children in different branches with the same parent node (third party), that especially ad networks result in longer tree branches, and that only 7% of all visited websites (TLD+1) never embedded a third party that might pose possible legal problems. Moreover, we have shown that studies that only measure landing pages of websites miss a substantial amount of embedded third parties and cookies set.

Third Party Dynamics We found that the challenges service providers face when they want to account for these challenges are due to the highly dynamic of the Web, not easy to implement. Accounting for all third parties present on each website loaded by users would create high overhead. Furthermore, the current legal uncertainty if service providers can be held accountable for the third parties embedded by their partners leaves potential risks that service providers face. Finally, we found that no one in the ecosys-

tem is directly to blame for the status quo, and solutions to the measured problems might result in substantial changes in the ecosystem. However, these changes—however they might look like—are necessary to offer users actionable tools to hold service providers and the embedded third parties accountable for the usage of users’ data, which at the same time can be implemented by the service providers on a technical level with reasonable overhead.

Cookie Syncing In contrast to previous work (see Section 2.3.2), we found statistically significant changes in the online advertising ecosystem around the GDPR enforcement date. Other work focused on embedded third parties [169] or, more specifically, tracking companies [41]. However, they could not measure the direct impact of the new legislation. Our results do not contradict other results as we also found that the ecosystem, in general, did not change. To a higher degree, our work shows that the effects of the GDPR might not be directly measurable in all aspects of the online ecosystem, but moreover, in-depth analysis is needed to get a better understanding of the effects of such complicated legislation in a complex environment.

CHAPTER 4

HUMAN ASPECTS OF THE GDPR

Based on our findings of the previously described measurement studies, this chapter takes a human-centric point of view on the usefulness and implementation of the *right to access* and *right to data portability*. We investigate how different companies implemented their data-sharing practices either by actively making use of the right to access, granted by the GDPR, or by expert interviews to understand the reasons why companies changed their practices in some areas (e.g., data sharing) but not in others (e.g., tracking). Notably, recent fines [78] and ongoing legal complaints [25]

for lack of transparency indicate that such aspects need to be studied in more detail. First, we analyze differences in the technical implementation of the new right (Section 4.1). We focus on the success and timing of requests, the data companies provide, and obstacles users face when they want to exercise their new rights. Second, we use the data received from the companies to test if users understand the data, and if they find it helpful to assess the privacy impact of a company. Furthermore, we interview companies active in the online advertising ecosystem to understand their point of view and to identify challenges they faced when they design processes that are compliant with the new law (see Section 4.2).

4.1 Analyzing SAR Implementations

The business models of modern websites often rely—directly or indirectly—on the collection of personal data. The majority of websites track visitors and collect data on their behavior for targeted advertising [49]. 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 [186]. As a result, the imbalance of power over information between data

processors (service providers) and data subjects (users) increased in the last years.

One of the GDPR's goals is to allow users to (re)gain control of the immaterial wealth of their 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 [177]. In this section, we make use of the new legislation and evaluate the subject access request processes of several companies. We identify prominent third parties on popular websites that collect tracking data and exercise our *right to access* and *right to data portability* 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 the 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 or used. Besides from the detailed overview of different approaches to implementing *subject access requests* (SARs) in practice; our work provides helpful pointers for companies, privacy advocates, and lawmakers how the GDPR and similar regulation could be improved.

4.1.1 The Rights to Access & Data Portability

Before we describe our analysis approach and the results, we give a short introduction of the new user rights most relevant in this section. Article 15 GDPR describes an individual's *right to access*. The right to access describes which information data controllers have to provide users upon request. This information includes data typically comprised in a privacy policy like a statement what categories of personal data are processed, the purpose of processing, or the right to complain to Data Protection Authorities (DPA). Besides, Article 15 GDPR grants users the right to access their data (*"The data subject shall have the right to obtain [...] access to the personal data [...]"*). Recital 63 describes the purpose of this access right for individuals *"to be aware of, and verify, the lawfulness of the processing"*. The recital also specifies that: *"Where possible, the controller should be able to provide remote access to a secure system which would provide the data subject with direct access to his or her personal data"*. In practice, a *subject access request* (SAR) is a way to exercise the right to access. Article 20 of the GDPR extends the access right to the *right to data portability*. Meaning that an individual may not only review data stored about them but can also request a copy of the collected personal data. Furthermore, the article defines that companies should provide this copy *"in a struc-*

tured, commonly used, machine-readable and interoperable format”.

Similar to the GDPR, the *California Consumer Privacy Act* (CCPA) [29] requires that starting in 2020 “*a business that receives a verifiable consumer request from a consumer to access personal information shall promptly take steps to disclose and deliver, free of charge to the consumer[...]*”. Furthermore, the CCPA requires that “*the information may be delivered by mail or electronically, and if provided electronically, the information shall be in a portable and, to the extent technically feasible, in a readily useable format that allows the consumer to transmit this information to another entity without hindrance.*” (Section 1798.100).

Verification of requests Legislators are aware of the problem of disclosing information to unauthorized individuals. The European and the Californian legislation both set standards on how an access request should be verified, to prevent fraud or unintended disclosure of data to unauthorized individuals, Recital 64 of the GDPR states that “*The controller should use all reasonable measures to verify the identity of a data subject who requests access, in particular in the context of online services and online identifiers*” while the CCPA is not (yet) that clear and only states that a measurement will be “*pursuant to regulations adopted*

by the Attorney General". While verifying access seems straight forward when there is something like a name or address to compare, it is unclear how companies can achieve the same verification when data is only stored pseudonymously and without having consent to collect/process it in the first place (see also Section 2.3.1).

4.1.2 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 ad data portability with these companies.

4.1.2.1 Approach

As the basis for this analysis, we use the measurement data and findings from our cookie syncing study presented in Section 3.3. Based on the collected information, we determine top companies that engage most in cookies syncing and top companies often embedded into websites. We chose 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 adequately informing users—which would make it quite hard for users to regain control of their 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 Section 4.1.2.3). Then we use our GDPR rights to access data portability to learn how companies respond to SARs and which data they provide to users.

4.1.2.2 Analysis Corpus

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*. Table 4.1 shows the companies included in our analysis corpus. In two cases (*Turn* and *BidSwitch*) parent organizations replied to our inquiries, in one case a subsidiary replied (*FreeWheel*) instead of the inquired company, and in one case (*RTL Group*) we were told to address our inquiry to a subsidiary (*SpotX*). Thus, our final corpus consists of 39 companies.

The first company that we did *not* include in the corpus (i.e., the 26th most embedded company) was embedded by

Table 4.1: The companies of our analysis corpus grouped by their respective business field(s). *AppNexus* (★) and *Adform* (♣) run two services and are therefore listed twice. *SpotX* is a subsidiary of the *RTL Group* (†).

Supply-Side Platforms			
Improve Digital Index Exchange	Smart AdServer	AppNexus★	Rubicon Project
	Sovrn		
Demand-Side Platforms			
TripleLift	MediaMath	Adform♣	AppNexus★
OpenX	DataXu	IponWeb (BidSwitch)	
Sizmek	Amobee (Turn)		
Advertising Companies			
The Trade Desk	Sharethrough	NeuStar	Criteo
Acxiom	SpotX†	Quantcast	
Data Management Platforms			
Lotame	Adform♣	Media Innovation Group	
Drawbridge			
Further Companies			
Google	Verizon	FreeWheel (Comcast)	
Oracle	Adobe	RTL Group†	Microsoft
comScore	Twitter	Harris Insights & Analytics	
Facebook	Amazon		

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

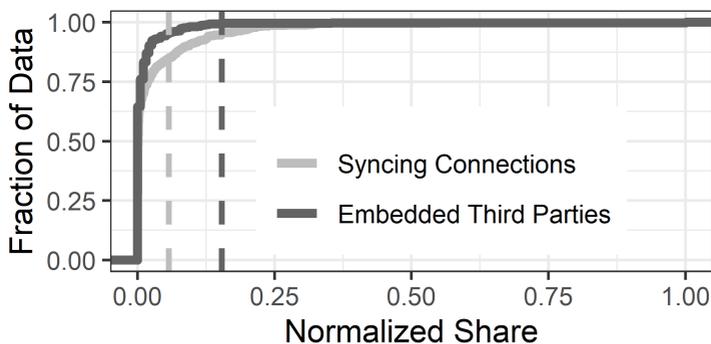


Figure 4.1: Shares of observed companies in terms of directly embedded (black) and actively syncing cookies (gray). The dotted lines show the share of the companies in our corpus.

in the corpus account for 66% of all ID syncing activities, while the remaining 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 distribution is also displayed in Figure 4.1.

The corpus consists of six SSPs, nine DSPs, seven companies that specialized in targeted ads, four DMPs,

and 13 companies whose primary business field not directly tied to the advertising. These remaining companies instead utilize ads to finance their services (e.g., *RTL Group*—a Luxembourg-based digital media group, *Facebook*—an online social network and media company, or *Verizon*—a telecommunications company).

While most of the companies in our corpus operate globally and run multiple offices, 82% have their headquarters located in the United States. The location of the remaining 18% of companies is in Europe (France, Luxembourg, the Netherlands, and the United Kingdom). This distribution is likely based on US/EU-based companies since we run our measurements from Europe. Since our goal is to measure the impact of the new European legislation, we expected this bias. We discuss the limitations of our analysis corpus in Section 4.3.1.

4.1.2.3 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. 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 for how long they store the data (retention time). 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 [41] that all services support and are therefore not listed. On the other hand, few services follow the “Do Not Track” signal, 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 [160]. 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 to exercise these rights. While the policies should also be specific on data sharing with third parties, companies are not required to list them individually but can describe them in categories.

4.1.2.4 Assessing the SAR Process

In order to test to which extent users can 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, companies need to disclose contact details of a person responsible for privacy-related questions (e.g., the Data Privacy Officer). 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.

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 conducting this survey to avoid biased responses. We evaluated the

response timings concerning 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.

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 data:

1. What information about a person associated with a cookie is stored and processed?
2. Where did the company get that information from? Did they receive it through further third parties?
3. Do they use the data to perform profiling?
4. With whom do they share information, through which channel, and which data precisely?

In round one, we used the following template to get in touch with the companies. During round two, we used similar sentences that contained the same information. We replaced the **bold** text with the information we extracted from our long term measurement. We pulled the corresponding cookies for each domain from our measurement logs and inserted the cookie IDs into the mails. As for the cookie syncing, we extracted up to five, chosen randomly

if necessary, instances of such syncing and inserted them into the mail.

Subject:

Request for personal data and additional information

Body:

Hi,

I noticed that there is a cookie stored in my browser associated with the domains **Domain 1** and **Domain 2**, which are owned by your company. Citing my European privacy rights, I would like to ask you to answer the following questions:

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?

I would also like to request a copy of the data kindly. The following cookies stored in my browser are associated with domains I found to be associated with your company:

- On domain: **Domain 1** key: 'id_key_1' and value: **personal_identifier_1**

- On domain: **Domain 2** key: `'id_key_2'` and value: `personal_identifier_2`

Another question I have is:

With whom do you share what information and how?

For example, I saw that you used the following IDs with your partners:

- partner: **Sync Partner 1** using the key: `'sync_key_1'` and value: `'sync_id_1'`
- partner: **Sync Partner 1** using the key: `'sync_key_2'` and value: `'sync_id_2'`

Thanks for your support,
Tobias Urban

In the inquiries, we used informal language such as that we did not quote any articles from the GDPR, nor did we use any legal terminology. We did so 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 and data portability. Actual users might have trouble accessing 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 [161]). We assume that a user who has privacy concerns can obtain this information, and usability improvements might follow soon.

4.1.3 Results and Evaluation

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

4.1.3.1 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 4.1.2.3. Table 4.2 provides a summary of the privacy policies of the companies in our dataset. It lists the most important tracking and GDPR-related attributes and what information they disclose. 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 shared. Only two of the policies disclosed and explain cookie syncing. Similarly, only eight policies 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 in which we were interested.

All policies, except for four, 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 [122]. 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 the 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 infor-

mation, 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.

Table 4.2: 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	unsp.	y	n.m.	n.m.	unsp.	7	account	n.m.	05/2018
Facebook	a,b,c,d,	unsp.	y	y	n.m.	differs	cat.	account	n.m.	04/2018
Amazon	n.m.	unsp.	n.m.	n.m.	n.m.	n.m.	cat.	n.m.	n.m.	08/2017
Verizon	a,b,c,f	unsp.	y	n.m.	n.m.	unsp.	329	website, email	n.m.	05/2018
App-Nexus	a,f	unsp.	y	n.m.	n.m.	3-60d, up to 18m	2309	website	n.m.	05/2018
Oracle	a,c,f	unsp.	y	health related	n.m.	12-18m	cat.	website	y	05/2018
Adobe	a,b,c,f	unsp.	y	n.m.	y	until opt-out	cat.	email, form	n	05/2018
Smart Ad-Server	a,f	unsp.	y	n.m.	y	1d-13m	cat.	email	n.m.	05/2018
RTL Group	a,c,f	unsp.	y	n.m.	n.m.	as long as necessary	cat.	email	n.m.	unclear

Company	Legal Basis*	Shared Data	3rd CO*	Sensitive Data	Profiling	Retention*	Partners*	Data Access*	DNT	Version
Improve Digital	a	listed	y	n.m.	n.m.	90d	cat.	email	y	05/2018
Media-Math	f	unsp.	y	health related ask to avoid	y	up to 13m	cat.	email	n.m.	05/2018
Triplelift	a,f	unsp.	y	n.m.	n.m.	as long as necessary	cat.	web-site	n	05/2018
Rubicon-Project	a,b,c,f	unsp.	y	n.m.	n.m.	90-366d	cat.	form	n.m.	05/2018
The Trade Desk	a,f	unsp.	US	not allowed	n.m.	18m, 3y aggregated	cat.	web-site	n.m.	10/2018
Share-Through	a,b,c,f	unsp.	y	n.m.	y	13m	cat.	email	n.m.	05/2018
Neustar	n.m.	IDs, segments	US	not allowed	n	13m + 18m aggregated	cat.	email	n.m.	08/2018
Draw-bridge	n.m.	IDs, segments	US	health related	n.m.	n.m.	cat.	email	n	08/2018
Adform	a,f	unsp.	y	not allowed	n.m.	13m	33	form/en	n.m.	unclear
Bidswitch	a,b,c,f	unsp.	y	n.m.	n.m.	“as long as necessary”	cat.	n.m.	n.m.	05/2018

Company	Legal Basis*	Shared Data	3rd CO*	Sensitive Data	Profiling	Retention*	Partners*	Data Access*	DNT	Version
Harris I & A	a,c	listed	y	y	n.m.	purpose fulfilled	cat.	email	n.m.	07/2018
Acxiom	a,f	cat.	y	no	n.m.	unsp.	cat.	register	n.m.	05/2018
IndexExchange	n.m.	aggregated only	US	no	no	13m	cat.	website	n	09/2018
Criteo	a	aggregated	y	no	n.m.	13m	61	email/m	n	05/2018
OpenX	a,f	unsp.	US	n.m.	n.m.	unsp.	cat.	email	y	05/2018
DataXU	a,b,c,f	behavioural	y	not in EU	n.m.	13m	cat.	email	n	06/2018
Lotame	n.m.	unsp.	US	health related	n.m.	13m	cat.	website	y	09/2018
FreeWheel	a,b,f	unsp.	Y	n.m.	n.m.	18m	cat.	email	n	05/2018
Amobee	a,f	unsp.	US	n.m.	y	13m	cat.	website	n.m.	06/2018
comScore	a,b,c,f	unsp.	y	n.m.	n.m.	n.m.	cat.	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.	unspecific	webform	n.m.	05/2018
Sizmek	a,b,c,f	segments	y	not knowingly	n.m.	13m	unsp.	website	mixed	05/2018

Company	Legal Basis*	Shared Data	3rd CO*	Sensitive Data	Profiling	Retention*	Partners*	Data Access*	DNT	Version
Twitter	a,b,c,f	listed	y	not allowed	n.m.	18m	16	account	n	05/2018
Microsoft	a,b,c,f	unsp.	y	y	n.m.	13m	>9	account	n	10/2018
Media Innovation	a	unsp.	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

4.1.3.2 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 to get access to the data associated with our cookie ID to evaluate the SAR process of each company, as described in Section 4.1.2.4. At the beginning of the first round (June 20, 2018), we sent out 32 emails and used six web forms to get in touch with each company. In the second round (September 21, 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 the law does not specify whether these refer to business or calendar days, we used two deadlines in our analysis (the dotted, gray lines in figures 4.2 and 4.3).

Response Types and Timing We grouped responses in three types: (1) *automatic* responses, (2) *mixed* responses, and (3) *human* responses. If the was identifiable as sent automatically by a computer system, we categorized it as “automatic” (e.g., a message from a ticket system stating that our request was received). 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. If there was any doubt, we ruled in favor of the companies. Figure 4.2 shows the number 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 round one, we received the most responses (51/100) during the first two weeks. We labeled the majority of these messages as “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 %

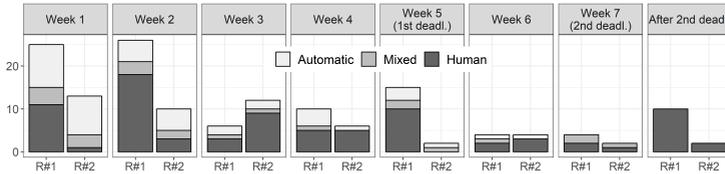


Figure 4.2: Types and timings of the received responses.

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). Only one company told us (in both rounds) that due to the complexity of our inquiry that they would need more time. In our second round of inquiries, we received fewer responses, approximately half of the amount. Partly, this is because we did not have to report any broken data access forms to the companies that we encountered in round one, which explains the fewer human responses in weeks one and two. However, we observed that in our second round, companies did not follow up on our further questions as they did in round one (e.g., if we asked for further clarification about data sharing).

Response Success The effort necessary to obtain access to personal data differed depending on the inquired com-

pany. To assess the workload of the process of a company, we use a simple scoring mechanism that mainly takes four factors with different impact into account:

1. amount of emails sent to the company before getting access to data associated with 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

This simple metric does 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 they perform profiling) that they did not answer 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 the paragraph “Subject Access Request Process” (see page 184), 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 4.3. 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 4.3a shows the results of our inquiries by the time of the first deadlines (July 20 and October 31, 2019). Note that it is unlikely that we provided a wrong cookie ID, but it is possible that companies do not have any data on the record because of short retention times or that they do not log some events, because there was no further interaction (e.g., the user did not click on an ad). Notably, 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 within 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, users could not know upfront that they needed this data 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. After the second deadline of round one, only 21 of 36 companies (54 %) shared data, or told us 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 correctly 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 4.3 shows that if companies granted access, we see that the workload is often quite low (in both rounds). In one case, with a high workload, in round one, we had a lengthy email exchange for getting access (in total, 13 emails—six sent by us). The other cases required a copy of the ID and, in one case, a signed affidavit. Notably, the overall workload in round 2 lowered, and companies usually wrapped up the process faster. The reduction of workload

Table 4.3: Overview of the SAR process and responses for both rounds of inquiries.

(a) 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 %

(b) 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

is because, on the one hand, we did not have to report broken SAR forms, and on the other hand, companies

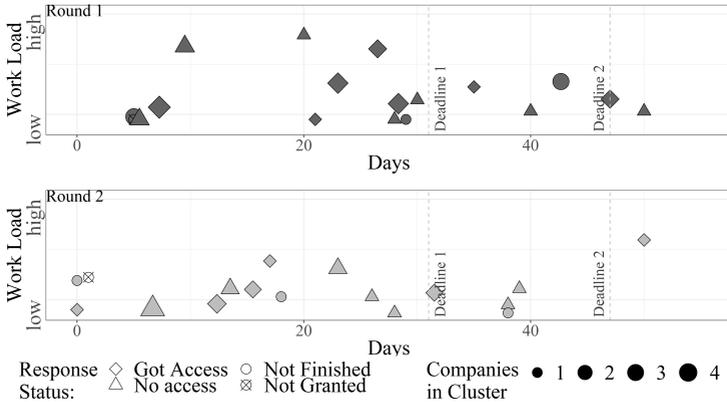


Figure 4.3: Comparison of the workload to get access to personal data companies stored about a user.

set up less “offline” obstacles. Notably, during round one, we observed that companies who claimed not to store any data still require multiple interactions before 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 Table 4.3b 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 that users 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.

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 data. If a company shared clickstream data (three in total), we manually checked if our dataset contained additional or missing websites that we had observed. In all three cases, the data was accurate. Overall, we grouped the received data 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. 4.5). *Tracking data*

is information on which websites the company tracked the user, also typically presented in a text file (see Fig. 4.6). *Segment data* is data companies inferred from a user's on-line behavior (see Fig. 4.4), 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 a company provided data they did not mention to collect in their privacy policy and many instances (all but one) where companies did not provide all data that might be collected.

Subject Access Request Process Companies handle inquiries very differently ranging from not responding at all, over directly 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 4.4a gives an overview of the obstacles users face when filing 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 when a user provides the cookie ID. 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 [64]. However, official interpretations state that data processors should have “*reasonable doubts*” before asking for additional data [13]. Those that request an ID card did not explain their doubt and did not describe how the ID helps

Table 4.4: Overview of the SAR process and responses for both rounds of inquiries.

(a) Obstacles			(b) Answers		
Status	R1	R2	Question	R1	R2
Affidavit	4	3	Q1 (data)	21	23
ID card	6	5	Q2 (sources)	6	6
Other	4	7	Q3 (profiling)	9	6
None	26	25	Q4 (sharing)	7	4

them to verify that the person requesting the data 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 Section 4.1.2.4). Only a few companies answered the additional questions we asked. Most of them referred to their privacy policy or did not provide further details.

Table 4.4b 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-

tion 4.3.1). Concerning Q1 and Q2, most answers contained references to or parts of the privacy policy.

As Table 4.2 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 company does not use the data for profiling. Six answers described in more detail how the data is processed and would suffice the GDPR rule to provide “*meaningful information about the logic involved*”. 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 shared data with specific companies 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 received a copy of their personal information.

Summary Our analysis of the implementations of the rights of access and data portability respectively highlighted the different interpretations of the GDPR. While the new

rights should enable users to “take back control” of their data, our results prompt the question of whether they can execute these rights. Furthermore, the usefulness of the provided data should be analyzed to test if it helps users to assess the privacy impact of a company.

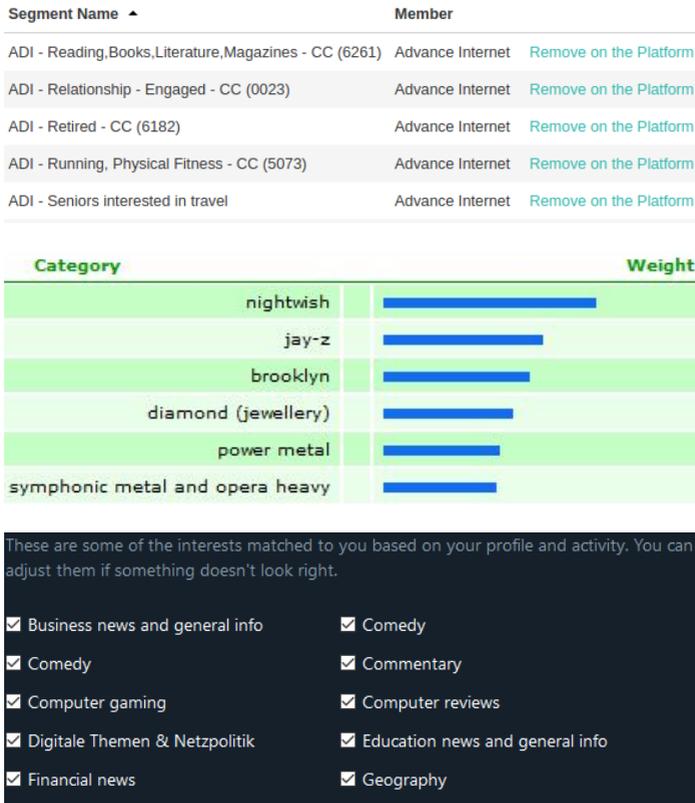


Figure 4.4: Inferred *interest segments* provided by different companies (anonymized).

```
{
  "cookieId": "DR-00000004,08871417",
  "ipAddresses": [ "161.185.160.XXX" ],
  "locations": [ {
    "postal": "07030",
    "city": "hoboken",
    "state": "new jersey",
    "country": "united states"
  } ]
}
```

timestamp	Dec. IP address	Internet Service Provider	Continent	Country	State	DMA ID	Numeric City ID	Postal code	Time zone	Browser ID	OS ID	Browser lang.
6/19/18 4:29	-1972356321	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:32	-1972356321	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:32	-1972356321	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:32	-1972356321	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:32	-1972356321	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:32	-1972356321	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:32	-1972356321	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:32	-1972356321	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:52	-1972356352	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:52	-1972356321	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:52	-1972356321	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en
6/19/18 4:52	-1972356352	Spectrum	us	us	nj	-99	7540	7030	-500	30	43	en

time	u_ip	ua	kvClob
24.09.18 09:29:25 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;cc=Country>st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3	
24.09.18 09:28:04 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;jpa=127.0.0.0/24;hb=1);dev=desktop;br=firefoxplnlinux_	
24.09.18 09:24:19 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;cc=Country>st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3	
24.09.18 09:24:18 UTC		Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cc=DE;st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3;	
24.09.18 09:17:29 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;cc=Country>st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3	
24.09.18 09:16:17 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;cc=Country>st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3	
24.09.18 09:16:17 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;cc=Country>st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3	
24.09.18 09:10:44 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;cc=Country>st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3	
24.09.18 09:09:36 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;cc=Country>st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3	
24.09.18 09:05:05 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;cc=Country>st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3	
24.09.18 08:58:43 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;cc=Country>st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3	
24.09.18 08:54:41 UTC	127.0.0.0	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; cs=;fst=-;cc=Country>st=state>dma=metrokey>l=city>pcc=zip>os=ISP>sp=3	

Figure 4.5: *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.ladepêche.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	https://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

Figure 4.6: *Tracking data* provided by different companies (anonymized).

4.2 Usefulness of Transparency Tools for Online Advertising

Based on the findings presented in the previous section, this section focuses on the usefulness of transparency tools and, more specifically, of the data provided by companies upon request. Previous studies have measured users' discomfort with ad personalization [116, 187], and highlighted the importance of transparency as a critical factor [46]. Therefore, scholars have argued that transparency is critical to counter the knowledge imbalance between tracking services and individuals that increments the discomfort [81].

To counter these problems, an increasing number of ad-tech companies offer ways to access such data via web portals (e.g., *TrippleLift's* approach [185]) or offer to answer data access requests via email. Through these means, users can gain insights into the data collected about them (e.g., sites they were tracked on) or information inferred from such data. New regulation fostered the industry's increase in transparency by new regulation introduced to account for the users' demand for more transparency [74]. The GDPR and the CCPA include the right of each user to request access to the data a company has collected about them (*Right to Access*, Article 15 GDPR). Prior

work on ad transparency only analyzed a small number of services offered by *Facebook* or *Google*, pioneers in this area [19]. The ongoing trend towards more transparency massively extends the number of services that have to disclose information and provide access to user data. In this section, we present a study on the extent of new transparency mechanisms and provide insights into how users and companies struggle with new opportunities and regulations.

In a system as complex as online advertising with multiple actors sharing and building upon tracking data, there are multiple challenges to *effective transparency* [143, 202, 115]. First, those collecting the data must be aware of what and whose data they directly or indirectly collect through third parties. Second, transparency is not an end in itself, so when companies provide personal data to data subjects, it has to be contextualized and presented in a way that conveys the essential facts but does not overwhelm the user. We study aspects of both challenges by evaluating the current state of transparency tools (see Section 4.2.1) and the data provided to users when they request access (see Section 4.2.2). While we first focus on the views and needs of users, we also try to understand the challenges companies face when providing transparency (see Section 4.2.3).

4.2.1 Analysis of Transparency Tools

Some ad-tech companies implemented ways to give individuals access to data stored about them to account for growing user demand. Most notable, *Google* and *Facebook* developed privacy dashboards or transparency portals after their data collection practices had come under public scrutiny [19]. Other businesses have also set up information sites or web forms that individuals can use to request their data. Recently, the number of available tools has grown to account for regulatory obligations of the GDPR and CCPA that require companies to give individuals access to collected personal data. Several examples of data such tools provide to users are shown in Figures 4.4, 4.5, and 4.6 (Section 4.1.3).

Companies to Analyze Previous work has shown how difficult it can be to get access to data and that requests not always successful [190]. To avoid the overhead of tedious and possibly unsuccessful access requests, we looked at all members of sizeable online advertising alliances (i.e., the *Network Advertising Initiative* (NAI), the *Interactive Advertising Bureau* (IAB), and the *Digital Advertising Alliance* (DAA)) and checked which company offered an online tool to access personal data. If a company offered an online tool, we analyzed it in our study. According to the public

statements of these alliances, they represent over 5,500 companies. However, they have only 500–600 (distinct) members listed on the organizations’ websites, which we all manually reviewed.

We analyzed all online tools we found and used the data we received from in the subject access request study (see Section 4.1). In total, we analyzed 22 web portals and responses to our subject access requests. We differentiate between two types of how users can access their data: *online* and *offline*. By *online* we mean that users can visit a website that (automatically) reads the user’s cookie store and shows personal data associated with the read identifier. If the data is provided in a file format (e.g., `.csv` or `.pdf` files), we labeled it *offline*. Using a VPN service (US–VA), we also verified that all online tools are also available to US-based IP addresses.

4.2.1.1 Criteria Definition

We evaluated the transparency tools based on heuristics from multiple sources: (1) user expectations elicited in previous studies, (2) descriptive information found in the privacy policies, and (3) self-regulative norms proposed by industry groups. In the following, we describe all the criteria we used to analyze the transparency tools of the 22 companies in detail.

User Expectations Previous work has shown that users have different—mostly negative—views on online behavioral advertising (OBA) but also demonstrate a need for transparency in OBA. In the following, we describe criteria found by previous work to be most relevant to users when it comes to transparency and understanding of OBA.

Interest segments/Demographics: Dolin et al. observed that users are more comfortable with OBA if they are aware of the connection between the created profile and their interests [46]. They also found that users' comfort with personalized ads is (positively) correlated with the perceived sensitivity of the data category (e.g., health-related information is critical). One approach to increase users' comfort could be to show the segment data assigned to the user by a company, although previous work has shown that these profiles tend to be inaccurate [148, 19]. Besides assumed interests, OBA is often based on (inferred) demographic information (e.g., ethnicity, age, location, or salary). Discrimination based on this demographic information is a big concern [145]. Displaying interest segments or demographic data to users can help them understand how companies use the collected data and why they see specific ads.

Tracking and Clickstream Data: Ur et al. found that users' views on online tracking can range from “useful” to “scary” [187], and other work highlights the dislike of

ads based on clickstream data [123]. Therefore, when companies disclose on what websites they have tracked users, it can be helpful to understand *why* certain ads are displayed, e.g., in cases of re-targeting where users see ads based on the products they have previously viewed online. As shown in Figure 4.6, clickstream data is made accessible in varying granularity ranging from raw data that includes user-agent and other technical information to lists of timestamps and websites recorded.

Chiasson et al. [31] analyze the willingness of users to share data with advertisers was analyzed. They found different factors that influence the willingness to share and show that users have complex privacy needs, which are not accounted for by current tools. Our work differs from the named approaches as we measure the *transparency needs* of users concerning OBA instead of analyzing their attitude towards different aspects of OBA.

To summarize, previous work found three criteria users expect to find when evaluating personal data an ad company collected: (1) interest segment data, (2) demographic data, and (3) tracking data. We inspected the data provided by the tools of the 22 companies and checked if the data can be grouped into any of these groups and also checked if the privacy policy stated if such data was collected/inferred.

Privacy Policies The content of privacy policies should be helpful to users who want to learn more about the privacy practices of a company, aside from being compliant with legal requirements. Therefore, we analyze the privacy policies of the 22 companies regarding three criteria: (C1) Does the policy use plain and unambiguous language?, (C2) Is the purpose of data collection explained?, and (C3) Is the way how data is collected explained? Thus, criteria C2 and C3 focus on the requirements for technical descriptions, not compliance in general.

To assess the readability of the privacy policies (C1), we used the Flesch-Kincaid Grade Level (FKG). This grading assesses how many years of school education one needs to understand a text (e.g., FKG = 12 means 12 years of education—senior-level high school students in the US). There is no consistent usage of readability scores in the literature, but Fabian et al. [55] found that the FKG, among other scores such as SMOG, RIX, or LIX, produces comparable results between each other (i.e., the correlation coefficients are almost 1). Previous work often does not report how they calculate the score (e.g., [55, 96, 95]), and we found that different available tools compute different FKG scores for the same text (e.g., because they compute sentence endings or syllables differently). We computed the FKG score using the *koRpus* R package [126].

Regarding criteria C2 and C3, the privacy policies of the companies were independently analyzed by two researchers with experience in the area to check whether they provide the required explanations. All researchers have a strong background in data protection, a strong understanding of the online ad ecosystem, and experience in the field. One is also a certified data protection officer with legal expertise. We told these researchers that a technical description, although it might not be understandable by most users, is sufficient. While this favors the companies, we assume that if users are interested in how or why data is collected, they could check what these technologies are and how they work. While it would be favorable, it is—from our point of view—not the purpose of a privacy policy to explain technical details of the used technologies.

Industry Self-Regulation The OBA industry associations have developed transparency guidelines for their members on how and which kind of information and choices they should provide to consumers (e.g., the DAA and IAB [43, 86]). As noted earlier, we analyzed the guidelines of the three most prominent alliances, the *Interactive Advertising Bureau* (IAB), the *Network Advertising Initiative* (NAI), and the *Digital Advertising Alliance* (DAA). All guidelines urge companies to take steps to increase trans-

parency to their users, and every company we analyzed is a member of at least one of the alliances.

However, the guidelines are quite vague and not easy to validate on the users' end. For example, companies should place a special icon on the ad if the company uses a behavioral profile to deliver it. The icon should provide a link to an opt-out tool and further explain why the ad is displayed. If the ad banner does not contain such an icon, this could mean that the company either provides an ad not based on a behavioral profile (e.g., depending on the site's content) or that it does not adhere to the self-imposed rules. By manual inspection of several websites, we found the same ad twice once labeled with the icon and the other without the label. It is possible that this observation was coincidental and that one ad was contextual, and the other was profile-based, but this illustrates that it is nearly impossible to decide if the guidelines were followed or not. Due to this inconsistency and previous work highlighting the ineffectiveness of such icons [156], we did not further investigate this transparency mechanism.

Besides the guidelines, the DAA, NAI, and IAB provide and maintain websites for consumers to learn more about online advertising and control privacy-related settings (e.g., mechanisms to opt-out of OBA for several ad companies at once). The DAA provides the *YourAd-Choices* [44] and the EDAA the *Your Online Choices* [54]

tool. At the time of this study, none of the guidelines contained rules or advice on how users can obtain access to their data.

4.2.1.2 Results

Table 4.5: Results of transparency tools analysis. ○: Does not apply. ●: Applies according to the privacy policy and data is provided. ◐: Applies according to the privacy policy, but no data is provided. ◑:= does not apply according to Privacy Policy, but data is provided. †: *Google* and *Facebook* only shows tracking data on their own platforms. *Twitter*'s way to provide sharing data did not work for us. *Sovrn* only shared pseudonyms of partners. ‡: Our analyzed profile did not include this data but could include it.

Company	Access		Expect. Segments Demographics Tracking	Privacy P.		Misc. Sharing explain tech.
	Online	Offline		FKG	explain why explain what	
Adform	✗	✓	● ◐ ●	13.40	✓ ✓	✗ ✓
Amobee	✓	✗	● ◐ ‡ ◐	13.05	✓ ✗	✗ ✗

Company	Access		Expect.	Privacy	P.		Misc.
	Online	Offline	Segments Demographics Tracking	FKG	explain why	explain what	Sharing explain tech.
AppNexus	✓	✗	● ● ●	11.96	✓	✓	✓
ComScore	✓	✗	● ● ●	10.63	✓	✓	✗
Conversant	✓	✓	● ● ^z ●	13.43	✗	✗	✓
Criteo	✗	✓	● ● ●	12.00	✓	✓	✓
Facebook	✓	✓	○ ● ● [†]	14.19	✓	✓	✗
Google	✓	✓	● ● ● [†]	13.37	✓	✓	✗
Leiki	✓	✗	● ○ ●	9.96	✗	✗	✗
Lotame	✓	✗	● ● ●	12.72	✓	✓	✗
MediaMath	✗	✓	● ● ●	12.69	✓	✓	✗
Oracle	✓	✗	● ● ●	11.92	✓	✗	✗
Quantcast	✗	✓	● ● ●	12.10	✓	✓	✗
Rubicon Project	✓	✗	○ ○ ●	12.77	✓	✓	✓
ShareThrough	✗	✓	● ● ●	12.07	✓	✓	✓
Sizmek	✓	✓	● ● ●	14.81	✗	✓	✗
Sojern	✗	✓	● ● ●	13.96	✓	✓	✗
Sovrn	✗	✓	● ● ●	14.24	✓	✗	✓ [†]
SpotXchange	✓	✗	○ ● ●	12.15	✗	✗	✓
The Trade Desk	✓	✗	● ● ^z ●	10.29	✓	✓	✓
TripleLift	✓	✗	● ● ^z ●	12.19	✓	✓	✓

Company	Access Online Offline	Expect. Segments Demographics Tracking	Privacy P. FKG explain why explain what	Misc. Sharing explain tech.
Twitter	✓ ✓	● ● ○	12.16 ✗ ✓	✗ [†] ✓

Table 4.5 lists the results of our analysis of transparency tools in alphabetic order. Ten companies provided data only online, eight companies provided data offline only, and five companies provided data in both ways. In general, online data can be seen as more usable because companies of often pre-processed it, while most ($n = 6$) of the offline data is comma-separated, which needs a more technical background to interpret. Some profiles contained inconclusive information (e.g., Segment: `companyB_Usersync_Global` or Your hashed IP address: `Ubuntu`). If we could see (or guess) the meaning of such information, we ruled in favor of the companies.

User Expectations We checked if the data provided by the transparency tools fits in these three categories, as described above. These categories contain information the ad companies collected inferred about the user regarding (1)

interest segments, (2) demographics, and (3) tracking. We inspected the data provided by the tools, checked if we can sort the data into any of these groups, and also reviewed if the privacy policy stated if such data was collected/inferred. We identified three different cases:

1. a company states that they collect data in one of the categories and provides this data (●),
2. a company states that they collect data on one of these categories but does not provide such information (◐), and
3. the company does not state that they collect data in one category and also does not provide any information (○).

We did not observe the case that a company did not state that they would collect data on a category but provided such information. In some cases, the 22 analyzed profiles did not contain interest segment or demographic data. However, the profiles might contain such data for other profiles because the shared data is not a full set of all categories but only the data they assigned to one user. We found four cases where this applies (marked with † in Table 4.5).

Thirteen companies provided information regarding inferred segments, and six companies chose not to provide this data, even though the privacy policy mentions that the company inferred segments. Nine companies provided de-

mographic information they inferred from the users' online actions. If the companies provided access to demographic information, most of them shared the user's location(s) or the inferred age. In general, users do not think that companies trying to learn their age is a severe privacy problem [145]. Two companies provided health-related information (e.g., *Health & Fitness > Diets & Nutrition*). No company provided racial information, which is in line with most privacy policies stating such information is not inferred.

While most companies state in their privacy policy that they track users around the web, we only found six companies that list the websites on which they tracked the user. *Google* only provided the visits to websites the company controls (e.g., *YouTube*), while the leading opinion is that they track users across tons of sites [49].

Legal Requirements Two researchers—with a strong understanding of the online ad ecosystem—were assigned the task to classify if companies disclose “why” and “how” data is collected. The final inter-rater agreement for classification shows substantial agreement (Cohen's $\kappa = 0.77$ for “why” and $\kappa = 0.74$ for “how” data is collected; agreement $> 90\%$ for both categories). We found that five companies do not disclose *why*, and five companies do not

disclose *how* they collect data. It is worth noticing that some companies only vaguely explain why they collect data (e.g., *<Company Name> “advertising technology allows Business Partners to target advertisements to users [...]”*) or how they implemented this. In three cases neither *why* nor *how* data is collected is given.

We computed the Flesch-Kincaid grade for all privacy policies. The score suggests that, on average, users need 12.58 (with SD 1.21) years of education (senior high school students in the US). Data provided by the *US Census Bureau* shows that 12% of all adult US citizens did not obtain a high school diploma [196].

Summary The field of analyzed transparency tools is heterogeneous concerning the type of data provided, the way of access, and information about sharing activities. Some companies reports’ lack of explanations of data collection and usage, and they do not provide all data representations that they claim to have. Besides, companies may not comply with the requirements of the GDPR regarding explanations of data collection and usage. As the tools provided data in different forms and levels of granularity, it is worth analyzing to which extent such data helps users to assess the privacy implications of a company’s activities.

4.2.2 Perception of Transparency Tools

Previous work has focused on the transparency of targeted ads themselves [156] or the accuracy of inferred-interest segments [19, 148]. In this study, we try to get a better understanding of the users' expectations and needs when it comes to transparency in online advertisement.

4.2.2.1 User Study Design

To get a better understanding of the users' side of transparency tools, we run an online survey which focuses on two aspects, First, we wanted to understand to what extent users can identify who is collecting their data because otherwise, they would not be able to request it. Second, the ways companies provide access to data differs from approaches studied in the past. Our goal was to understand how different types of data disclosures found in the field help participants understand the privacy implications of a company. Our study focuses on the ability of users to understand the provided data as it the most important mechanism to provide transparency. At the same time, other aspects like completeness or the comprehensibility of how companies inferred specific information also play an essential role in the value of these tools.

To test if users can identify which ad network is responsible for an advertisement, we present two screenshots of websites, one of which contained a standard ad banner and the other an advertisement with links to articles distinguished as “Recommendations”. The recommendation contains a reference to the third party that generated the ad (i.e., “*Recommended by Outbrain*”). The ad banner contains a link to an opt-out program but does not directly show the name of the third party. Users would have to hover over the ad with the mouse and check the URL displayed in the browser’s status bar to identify the ad network—we included this URL in the screenshot but did *not* highlight it.

The remainder of the survey focuses on the users’ expectations and understanding of personal data provided by ad companies upon request (i.e., *subject access request*). To assess this, we took screenshots of nine real-world profiles that we received upon request, as explained previously (see Section 4.2.2.1).

We grouped the nine profiles into three categories based on their content: (1) technical data, (2) tracking data, and (3) segment data—we define these categories in Section 4.2.2.2. The order of these categories was randomized for each participant. To reduce the length of the survey, we did not differentiate between segment and demographic data. Plane et al. studied the influence of this data on

The screenshot shows a web browser displaying an ESPN article. The browser's address bar shows the URL `www.espn.com/mma/story/_/id/25773451/`. Below the browser, there is a navigation bar with the ESPN logo and a menu of sports events including NBA, Final TOR BOS, FinalOT BKN HOU, Final MIL MEM, Final SA DAL, Final CLE POR, and Final N G. The main content area features a quote from Joe Scarnic/Getty Images: "These are the goals that I set out since I was a little kid: to be an Olympic champion and to be a UFC champion. When people talk about being a double champ, the true definition of a champ-champ is Henry Cejudo. An Olympic gold medalist and a UFC champ. Being a champ-champ in the UFC in two divisions is a two-division world champ. I'm the champ in two different sports, and I think that's the difference between me and my opponent this Saturday." Below the quote is a red-bordered box containing a "SPONSORED HEADLINES" section. This section is titled "Recommended by Outbrain" and features three articles: "This Incredible Trick Protects Your Computer For Free" by The Review Experts, "Where the World's Billionaires Live" by Mansion Global, and "This \$99 drone might be the most amazing invention in 2019" by ZaboTech.

Figure 4.7: Article recommendation including the company providing it (top right corner).

the perception of online ads before [145]. Instead, we differentiated between more abstract clickstream (tracking) and detailed technical data. Our analysis of existing trans-

The image shows a screenshot of a Reddit page. The browser address bar displays the URL: <https://www.reddit.com/r/gameofthrones/comments/afqsjj/s/>. The page content includes a search bar, a login/sign up prompt, and a list of comments. The comments are as follows:

- TheUnknown285** (3.7k points, 2 days ago): "Well, we finally got Catelyn as a stoneheart, just not how we were expecting." (1 reply)
- starkey2** (576 points, 2 days ago): "The quote was a little rough on her; that was when she was at her worst." (1 reply)
- NewZealandTemp** (428 points, 2 days ago): "Admitting her faults almost makes it easy to forgive her for treating Jon that way. She has to live with what she sees as her and her husbands biggest shame every day." (1 reply)

Below the comments, there are links for "26 more replies" and "42 more replies". On the right side of the page, there is a red-bordered advertisement banner for **THE OUTNET.COM**. The banner features a grid of six fashion items: a peach-colored dress, a beige top with a floral brooch, a black high-heeled shoe, a black and white patterned jacket, a blue and white sneaker, and a black high-heeled boot. The text "ADVERTISEMENT" is at the top left of the banner, and "THE OUTNET.COM" is at the bottom. A mouse cursor is visible over the black high-heeled shoe.

At the bottom of the browser window, the URL https://www.googleadservices.com/pagead/aclk?sa=L&ai=Caz4Epy...YCh3JUgWXEAEYASADEgLTvD_BwE&client=ca-pub-9699191536922730 is visible.

Figure 4.8: Traditional ad banner.

parency tools (see Section 4.2.1) showed that disclosing data in this form is standard. The new right to data portability also brought the disclosure of raw data. The perception of this form of data for transparency was not studied yet.

Each section of categories starts with a brief introduction to the data shown, followed by three different examples

of profiles. Participants had to assess four statements regarding their understanding and the presentation of the data. We randomized the order of the profiles within each category. At the end of each section of categories, we asked the participants if these profiles would help them to assess the privacy impact of the companies better. Following each section of categories, we asked participants general questions regarding their personal views on provided data and preferences, which data representation and the category they prefer. For most questions, we used 5-point Likert scales, and, for the remainder, we used “Yes/No” questions and a prioritization question. In our survey, all questions (aside from optional open-ended questions) provide a “*I prefer not to answer.*” answer option.

In general, we used Pearson’s chi-squared test to test the independence of two variables and the Pearson correlation coefficient to determine a linear correlation between two variables. For both tests, we used a significance level of $\alpha = 0.5$. Furthermore, we assigned the value 5 to the most positive answer on a Likert scale, 1 to the most negative answer option, and consequentially three to the natural option. (e.g., “Strongly Agree” = 5, “Undecided” = 3, and “Strongly Disagree” = 1).

We conducted a pre-study ($n = 50$) with a similar survey structure, as described above. In the pre-study, we focused on how users might use data access to their benefit

(i.e., how they would use the provided information). However, due to usability problems, users did not give useful feedback on this (e.g., P-6 stated: “*No, this is gibberish to me*”). Thus, we dropped this question for the final study. The full questionnaire can be found in Appendix A.1.

To recruit participants, we used Amazon’s *Mechanical Turk* (MTurk) [7] and only accepted participants with high task completion rates ($\geq 97\%$) and permanent residents of the US. Furthermore, we only accepted participants who were at least 18 years old and asked for their consent to participate in the survey. In the introduction of the survey, we disclosed our names, affiliations, and all sponsors. We used a self-hosted LimeSurvey [114] instance to conduct the survey. Participants received \$3 dollars for completing the survey, which took them around 15 minutes on average (median = 13 min). We saved all answers pseudonymously using a unique random string, used by MTurk to pay the workers. After the payment process with MTurk had been completed, we deleted the identifier to increase participants’ anonymity level.

4.2.2.2 Results

In February 2019, we surveyed 490 participants (using Amazon’s Mechanical Turk). In the following, we describe the main results.

Participant Demographics 54% of the participants are male, the majority (46%) of participants is between 25 and 34 years old, and holds either a high school diploma (33%) or a bachelor's degree (52%). The full demographic information in our study, compared to the general adult US population provided by the *US Census Bureau* [196], can be found in Table 4.7. Our sample is biased as more participants identified as males, have a better formal education, and are younger than the general population. A majority of participants (90%) use some form of privacy protection online, at least from time to time. Table 4.6 shows the used mechanisms. The number of adblocker usage is higher than the previously reported [48]. 50% of participants reported that they use an ad blocker. This number is higher than previously reported. A recent study found that 37% of Internet users use an ad blocker, especially younger individuals [48]. We assume that we observed more adblocker usage since our sample is skewed towards younger participants.

Attitude Towards Online Advertising The general view of participants on online advertising is quite neutral. Participants who see ads that suit their interests still evaluate them slightly negatively (mean: 2.7 with SD 0.4 and the hypothesis test yielded $p < 0.0005$), which is in line

Table 4.6: Privacy protection methods used by participants. Each participant used at least one the listed method or is listed among “None”.

Adblocker	Private Browsing.	Delete Cookie	Opt-Out	None
50.4%	51.8%	70.6%	31.2%	10.0%

with the previous work [46] (Q1 and Q2). Other studies also found that users find ads “creepy” or “intrusive” [187]. In our study users expressed such views too, but at the same time they did not evaluate ads negatively (e.g., P-02 stated in Q9: “*They [ads] are creepy, a product is merely mentioned in my house then I see ads for it the next time I’m online*” but at the same time stated her views on ads as “Moderately satisfied”).

The neutral view on online ads is likewise observed in an open-ended question (Q9) and on a Likert scale. 85% of participants choose one the following answer options almost balanced (Q2): “Moderately satisfied”, “Neither satisfied nor dissatisfied”, or “Moderately dissatisfied” with a mean of 3.07 and SD 0.5. In total, 73% of participants “*agreed*” (47%) or “*strongly agreed*” (26%) with the statement that access to personal data is useful to assess privacy implications of the usage of their data (Q5), but only 19 par-

Table 4.7: Participant Demographics. One participant preferred not to answer demographic questions (sex and age) and 4 preferred not to state their education level. ★: The census data does not account for non-binary individuals. ♣: The census data combines these categories.

	Amount	%	US pop.
Gender			
Male	264	54 %	49 %
Female	224	46 %	51 %
Non-binary	1	0 %	— ★
Age			
18–24	41	8 %	16 %
25–34	218	46 %	22 %
35–44	112	23 %	20 %
45–54	66	14 %	21 %
55–64	40	8 %	21 %
Education			
None	1	0 %	12 %
High School	161	33 %	51 %
Bachelor’s	255	52 %	18 %
Pro./Master’s/Ph.D.	69	14 %	11 % ♣

ticipants requested their data, mostly from big companies like *Google* or *Facebook* (Q4).

In Q6, 60% of participants stated they were “not” (10%) or “*somewhat knowledgeable*” (50%) about online advertisements while 30% stated that they were “*very knowledgeable*” (5%) or “*knowledgeable*” (25%). We used a multiple-choice question to test this self-assessment (Q7). This question contained seven statements on the online advertising industry—four of which are correct and three incorrect. Each (multiple-choice) answer option was similarly often selected (mean: 357 times with SD 51). Furthermore, participants stated misconceptions in an open-ended question (e.g., P-152: “*I honestly expect some ad companies to illegally collect my facial expressions and sounds in my environment through cameras and microphones. I always expect them to access other apps and histories of everything that I do*”, while there are no reports on this happening in practice. However, the most common misconceptions were that ad companies have access to *all* purchased products (64%) and the *full* browsing history (66%).

65% of all participants stated they thought that companies did *not* provide all collected information upon request, only 13% thought they would do that, and 22% had no opinion on this topic (Q22). This mistrust might be based on misconceptions about what companies can collect or relate to public reporting on data leakage scandals (e.g., *Facebook* providing data to *Cambridge Analytica* and *not* informing users properly [178]).

Identifying Data Collectors We aimed to assess if users understand which personal data companies used when they provide ads to users and if users know whom they have to ask to get access to this data (Q10 and Q11). To do so, we showed users two screenshots of websites containing ads (see Figure 4.7 and Figure 4.8); one contained information regarding the company providing the ad, while the other did not. We asked users whom they would have to contact if they wanted to understand *why* they see this specific ad. We provided different answer options (multiple choice): (1) the website on which the ad appeared, (2) the company name of the advertised product, and (3) the ad network providing the ad.

In the case of the ad that contained the ad network's name, 46% of users answered the question correctly, 28% named the advertised company, 17% named the visited website, and the remaining 9% did not know whom they have to ask (Q11). For the ad banner that did *not* directly include the ad network's name but showed it in the link when hovering over the ad (Q10), 43% named the advertised company as the contact company, 24% named the visited website, and only 24% correctly knew whom they should have to contact. In conclusion, only a minority of participants were able to identify the right company to contact, but we did not find a significant correlation between users' self-assessed knowledge about online advertising and

their answers (Pearson-correlation: $r = 0.56$ with $p = 0.57$). Hence, users have trouble identifying who is collecting their data when it comes to ads, but stating the name of the collecting company close to the banner helps.

Assessment of Provided Data In the survey, participants were shown three data category sections, each listing different data types we received by using the different transparency tools (Q12–Q21). The categories are (1) technical data, (2) tracking data, and (3) segment data. *Technical data* is raw data presented to users often without any pre-processing, typically in a text file. This type of data is likely log-data that can be directly extracted from HTTP traffic (e.g., user agents) or directly derived from this data (e.g., locations based on IP addresses). *Tracking data*, also often provided in text files, is information on websites on which the ad company has tracked the user (clickstream data) or with whom personal data was shared. Interest *segment data* is information ad companies inferred from a user’s online behavior.

Figure 4.9 highlights the results of participants’ views on the provided data. Tools that shared inferred data (i.e., segments) were evaluated very positively in all four question categories. Over three quarters (76%) of participants stated that they understood the provided data and

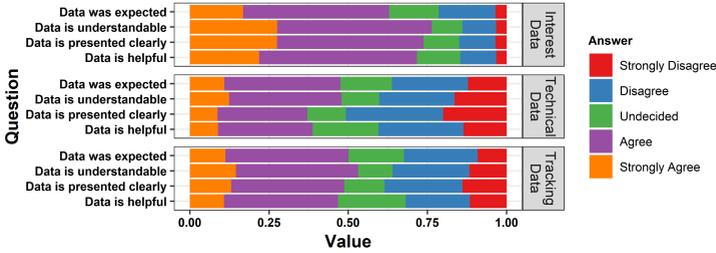


Figure 4.9: Evaluation of different aspects of the provided data.

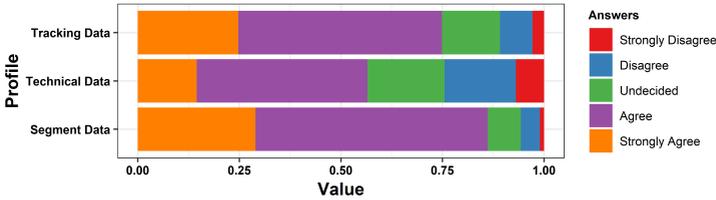


Figure 4.10: Participants' view on the usefulness of profile categories.

thought that it was helpful to understand what companies do with personal data (sum of answer options “*Strongly Agree*” and “*Agree*”). It is worth noting that not all companies infer this kind of data. For example, a company offering a service to identify ad fraud will most likely not

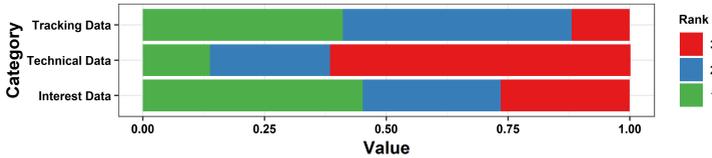


Figure 4.11: Participants' ranking of profiles.

infer high-level data but still has access to personal data like IP addresses. Participants understand tracking information less easily: More than half of the participants (53 %) report that they understood the data and found it helpful (47 %). Profiles that provide technical data were rated slightly less understandable but much less helpful (39 %), and these profiles present data in a much less clear way (37 %) than in the other categories of profiles. We found a correlation between all four questions on clear presentation, understandability, helpfulness, and whether users anticipated this type of data, in each section of categories (Pearson correlation: $r > 0.5$, $p < 0.001$).

After presenting all three profiles of one category, we asked participants if such data is useful to assess the privacy impact of companies. The results are given in Figure 4.10. Similar to the assessment of profile categories, segment data is rated to be most helpful, followed by tracking data. When it comes to preferences which data users would like

to receive when they performed an access request (Q23), participants equally rank “tracking data” and “interest data” as first choice (41 % for tracking and 45 % for segment data) but more users chose “tracking data” as their second choice (47 % vs. 28 %—see Figure 4.11). This discrepancy is unexpected as participants stated they found segment data to be most useful (see Figure 4.10) to assess the privacy implications of a company, and one would expect that they also 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 ($p < 0.0001$) and stated that the presentations were transparent ($p < 0.0001$). Furthermore, participants who stated that the presented data was useful—regarding the usage of data—also stated that the data helped to bring more transparency to the advertisement ecosystem ($p \approx 0.0005$). Access to technical data is by far the least favorite way how users want to get access to their data. However, technically-savvy users who answered our question regarding the ecosystem (Q7) correctly and stated that they were at least “*knowledgeable*” (Q6) about the online tracking industry rated this kind of information more helpful ($p < 0.001$).

When asked (Q24), 55 % of participants expressed that, after seeing personal data collected on a stranger, (i.e., data displayed in previous questions was based on one of the

author's access requests) 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 (Q25). 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.

Summary The analysis of users' perception of available transparency tools shows that individuals prefer interest segments over raw technical data when they get access to collected data. Furthermore, while participants were aware of the privacy impact of ad companies, only very few requested access to their data but reported they might do so.

4.2.3 Business Perspective

Aside from evaluating existing transparency tools, we contacted 773 companies to assess their view on the usefulness of transparency tools for users. We motivated our inquiries not purely on transparency tools in general but also wanted to assess how current legislation (i.e., the GDPR) influenced the design of such tools and which challenges companies faced designing these tools. All of these companies participate in at least one of the “opt-out tools”

(e.g., The DAA’s tool [44]) provided by different ad alliances. These companies are comparable to the members named on the alliances’ websites, and therefore we think that these companies are a representative set of companies participating in the online advertisement ecosystem. To this end, we crawled 773 privacy policies of the companies we analyzed and extracted all email addresses present in the policies, using a regular expression. By manual inspection, we identified the relevant email addresses by name (e.g., `privacy@company.com`) and dropped other addresses (e.g., `sales@company.com`) if we found a more specific one. We retrieved the email addresses from the privacy policy of each company and invited them, via email, to participate in an online survey. We did not find an email address in 35 policies via automatic or manual inspection. We used the regular email server from a research organization to send the emails.

4.2.3.1 Company Study Design

We sent the first batch of email invitations to 74 companies (approx. 10 %) in October 2018. In December 2018, we invited 333 companies (approx. 45 %) to participate in our survey. In January 2019, the final batch of 333 companies was invited. Out of all invitations, 147 resulted in an error message. Forty-eight companies classified our email

as spam or batch mail, and, therefore, they dismissed it. The rest resulted in other delivery errors. We manually double-checked all of these email addresses for transcription errors. While one company did not exist anymore, the other companies listed email addresses in their privacy policies that do not exist at all. Overall, 593 companies received an invitation to participate in our study.

The survey consists of four categories of questions: (1) general questions regarding the companies business processes for “*subject access requests*” (SAR), (2) four questions regarding their development of SAR processes, (3) three open-ended questions regarding the company’s view on the usefulness of data transparency and the GDPR in general, and (4) demographic questions. The full questionnaire can be found in Appendix A.2. At the end of the questionnaire, we asked participants to volunteer for an in-person interview.

We used a semi-structured guideline as described by Flick [59] to conduct the interviews. After a brief introduction, interviewees were asked to explain their job position, if applicable, the business model of the company they are working for, and then give a broad assessment of what the GDPR meant for their business. Afterward, the interviews focused on the same topics as the survey, but we also asked participants for their personal opinion about how to improve transparency. We conducted the interviews in

January and February 2019. All companies that completed the survey and participated in the interviews did so on the request of anonymity and are not necessarily the companies whose privacy tools we analyzed in Section 4.2.1. We transcribed the interviews focusing on content. The transcriptions were analyzed by first paraphrasing their content, identifying topics, and comparing the different interviewees' position on the identified topics. In two cases, the analysis is based on notes because one participant did not agree to be recorded, and in one case, the recording was of bad quality.

4.2.3.2 Results

Of the 593 invited companies, 24 (4%) completed the survey, and eight agreed to participate in an in-person telephone interview. In 14 cases, the company's Data Protection Officer or a person from the legal department took the survey. In the remaining cases, we got responses from persons on the executive level in the company (e.g., "head of data business" or CEO). Twelve of the participating companies have their headquarter within the EU, nine outside the EU (eight in the US and one in Israel), and two companies preferred not to disclose this information. In nine surveys, the companies reported that they employ 1–100 people, ten had 101–500, and four companies had

more than 1,000 employees. The size of the department responsible for handling privacy-related requests is in all but two cases proportional to the company size (i.e., larger companies have larger privacy departments). The companies where this is disproportional are law firms consulting other companies regarding privacy-related topics (e.g., GDPR compliance).

Of the eight interview partners, four companies directly collect and process personal data, two companies only handle personal data on a business-to-business side, and two companies handle both. All stated that they handle personal data of at least one million data subjects. Two of the interviewees were privacy consultants and claimed to work for “dozens” of clients in the ad industry. In the interviews, they answered the questions generalized for all of their clients. We report all findings and quotes in the singular to provide the same level of anonymity to all interview partners.

Impact of the GDPR In the interviews, several partners highlighted that—from their point of view—achieving GDPR compliance in online advertisement is one of the hardest tasks (I-2: “*So I represented advertising in the GDPR groups, so we kind of created a group of champions, if you want, so I was the champion on advertising, and it proved*”).

*that advertising was the most challenging one [...]”). Not surprisingly, all participants stated (in the interviews as well as the survey) that the GDPR helped them to convince their management to invest in privacy. However, aside from these increased costs, the only other benefit mentioned by more than one company was that GDPR compliance might be a competitive factor (e.g., P-17 stated: “*It became a marketing talking point that we used to show that people want some control over what data is collected about them, but other than that it has had almost zero benefits and quite a bit of cost*”).*

When asked about the impact of the GDPR on their daily work, we got mixed responses. Some companies stated—aside the two months before the GDPR—that the workflows in their companies not changed much while others reported problems that, as a result of the GDPR, data they had considered non-PII now is considered PII data (e.g., I-4: “[...] *suddenly, any identifier [sic.] would become PII*”). As described in Section 2.1, the GDPR has a broader definition of personal data that not only includes directly identifiable information. “PII” is the term used in the legal debate in the US. Interview partners that claimed their processes had already been GDPR complaint beforehand stated that they had to add modules that implement the required accountability (e.g., records when and how consent was conceived). One company stated that they

had eliminated all personal identifiers not necessary for their core business, a data minimization approach favored by privacy scholars. One of the consulting companies mentioned that they doubled the number of their clients during the month before the GDPR went into effect.

Access Requests We asked survey participants about subject access requests since the enforcement of the GDPR. While 11 companies stated that they had a standardized access request process (e.g., a website that presents the collected data to the user), slightly more (13) stated that they handled each SAR individually (e.g., the data is pulled from a database by the DPO). Responses show that overall fewer individuals than expected requested access. Twelve participants stated that they received the expected amount, and ten stated that they got “less” (4) or “way less” (6) than expected (one company preferred not to disclose). One interview partner (I-7) reported that they and other companies had expected “a five-digit figure of requests per month”, but over the whole year, they had received less than ten access requests. Consequently, half of the companies (12) do not consider changing how they handle access requests; only six of the participants reported that they were planning to change the process. The remaining

interviewees said they would like to change the process or reported that they already changed it.

Regarding guidelines, 15 companies stated that they preferred more unified guidelines about how to handle access requests. When asked who should provide this guidance, nine participants would prefer self-regulation (e.g., by the IAB), ten normative guidance (e.g., ISO standard), and three legislative guidance (e.g., amendments of the GDPR). These rather high numbers are likely related to the uncertainties that come with the new legislation. In the interviews, mainly representatives from smaller companies said they had a hard time defining processes they thought would comply with the new legislation. Thus, the desire for more regulation or guidelines could be a result of frustration companies faced when trying to comply with such complex regulations.

Challenges In the online survey, we asked participants about the challenges for transparency and access requests. We grouped the answers to the questions into two categories: (1) identification of users and (2) uncertainty about dataflows. One of the main challenges for companies is to decide what level of identification they required before answering access requests. While some rely on the cookie ID, others require a signed affidavit. On the one hand, it

is necessary to make sure that the company only grants access to the actual individual whose data is stored. On the other hand, “*authentication of individuals creates more sensitive data*” (P-16).

The challenge of identification was also highlighted during the interviews. All but one of the participants agreed that identification is problematic since a cookie ID is not sufficient to identify a person. One company stated that they were not concerned about cookie ID misuse, and all other companies implement the process differently, ranging from signed affidavits to screenshots of browser cookie stores. In terms of provided data, all companies claimed that they provide all data upon request—showing the apparent discrepancy between user perception, as stated in our survey, and company claims. The second challenge is related to the ad ecosystem, which depends on data exchange between companies. In order to provide access to data, services have to “*understand the dataflow of each partner [they] work with*” (P-13). While some companies most likely created a record of processing activities as required by the GDPR, one of the interview partners claimed that not all are aware of all dataflows: “[...] *one of the reasons is technical because I’m quite sure that they do not know how much data they have. I’m positive on this.*” (I-1).

Opinions In the last part of the survey and interviews, we asked participants about their opinion of the GDPR and the new transparency rights. They assessed that the GDPR brought more attention to privacy in terms of management. This increase is mostly because of the high fines, but also because of public awareness. Therefore, all agreed that privacy has become more integrated into daily practices. If participants had an opinion on why users have such a negative view of ad-tech companies, they said that the main reason was the general negative attitude towards advertisements and data leaks was the main reason for that. One participant was puzzled by our finding that 65% of users believe that companies do not share all data upon request and stated that from his/her point of view this was not correct.

Regarding transparency, all participants agreed that this was a positive development but added that the current systems provide only a few benefits to users and are often impractical in practice. As I-3 put it, “*it is not useless, but users—and that is why they are not contacting us—that do not want to be tracked already have an app [to block tracking]*”, while other stated that SARs and privacy policies are “*sufficient*” (I-1) measures. At the end of the interviews, we asked what the interviewees thought would help to bring more transparency to the online ad industry. Three interview participants described a system that

would give users high-level information about types of data collected and which would summarize them similarly to credit scores, while still offering the technical data to those who are interested. Furthermore, two partners highlighted that companies should address transparency when the data is collected or shared and that SARs do not help with that. Another idea was to add standardized ‘traffic sign’ like symbols to websites, ad banners, or cookie banners that easily describe which data is collected and for what purpose. When asked why companies format the provided data in non-usable ways, partners named different reasons like the data would “*become stultified*” (I-1) if raw data was formatted in a (standardized) way. One participant said that they could not do that because then they would become a data processor and not only a “*data collector*” (I-5). Another participant was surprised that not all companies provided inferred data like segments: “*Then you have been given an incomplete dataset because that is the only way that it actually works well for an ad company is to have a better understanding of what your preferences are. So I think then it is time for a complaint.*” (I-6).

Summary Our study of the company perspective on data transparency did not reveal a consensus across topics. While some would prefer more detailed guidelines, others

think the current implementation is sufficient. Similarly, approaches to subject access requests vary, and there appears to be little support for improvements as the main objective is legal compliance instead of fostering understanding at the users' end. While most participants think there is value in transparency mechanisms, there still seems to be a mismatch between user expectations and the understanding of personal data in the industry that has not considered users' data personal data for so long and does not always seem aware of the actual dataflows.

4.3 Discussion

The GDPR puts ad tech companies in quite a dilemma when it comes to granting users access to their data. On the one hand, they want to make it as easy as possible for users to exercise their right to access (e.g., by providing a website that automatically displays all information based on the cookies set in the browser). On the other hand, they want to prevent misuse as some companies might leak long browsing histories of some users since most keep the data for over a year. There is a general mismatch between the definition of personal data about one data subject and the socio-technical context. For example, it is not able to model use cases in which multiple persons use a device. These services cannot check who used a device and left the possibility that someone gets access to data that another user of the same device might have intentionally deleted.

The companies we surveyed reported that the number of users asking for access is rather low. One factor for this is probably that many consumers are not (yet) aware of these new opportunities, granted by the online tools and new regulation. Still, some companies make it hard for users to learn who collects their data so that even those who know about the new tools or their rights might have a hard time executing them. Still, our user study shows that companies are making it complicated to know who is

collecting their data so that even those that know about their rights might have a hard time executing them.

When getting access, users prefer receiving inferred information (e.g., interest segments) rather than technical data. In our user study, participants strongly expressed their wish that the provided data should be “*less technical*” (P-43) and in a “*easy to read visualization*” (P-324). However, not all companies can provide such information because it depends on the business the company is doing. New ways to display technical data in a meaningful way are necessary here. Furthermore, companies should provide both high-level information that users can easily understand and the underlying raw data so that users can use a tool to analyze the data according to their own needs. The problems with privacy policies have been long known, but still, it seems to be up to research to develop better tools, as companies do not focus on understanding but more on legal compliance.

Over 60% of the companies that participated in our survey expressed the wish for more regulatory guidance on the design of access request processes (i.e., transparency tools). Our research suggests that consumers would appreciate simplified and unified ways to obtain transparency. Based on our results, we recommend providing a visual overview (e.g., a workflow diagram) that describes what happens with the collected data, where it comes from, and

with whom it is shared. Further research in this area is needed to evaluate designs. An industry-wide standard would allow users to compare two services regarding their privacy impact. Those that want to be transparent about their practices should start by educating users about what they do with personal data *before* collecting and presenting it, first on a high level with the option of downloading raw data.

It is the public mistrust in the data sharing industry that fueled harsher regulations. To counter misconceptions, companies need to improve public understanding of their practices. It could help to provide information on what is *not* done with collected data (e.g., a company might not collect the users' location but users might still expect to find it). As we found that users struggle to identify the companies that might collect personal data, it would be helpful to add "Provided by *X*" information in every ad banner.

4.3.1 Limitations and Ethical Considerations

In the following, we discuss the limitations and the ethical considerations we took in our human-centric analysis of the right to access and right to data portability.

Implementation of Subject Access Requests In our analysis of the SAR process, 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 should be validated with user experiments. Right now, it serves only as an approximation. We cannot check whether or not the companies answered truthfully and provided all the data they stored, shared, and processed. To check that we would need to have direct access to the services' databases. For the same reason, we were also not able to measure what information companies exchanges on separate channels besides the synced IDs.

Our research on SAR processes (Section 4.1) includes human subjects (the persons exercising their rights and the persons responding to our requests), and therefore we took ethical considerations 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. We did not choose to debrief most of the companies since we exercise our rights granted by

the GDPR, and we do not study the persons replying, but the process of how companies handle the new regulations. Note that we contacted the companies that did not respond or had a poorly designed process, without any responses. When contacting the companies, we did not disclose we conduct a scientific survey, but we disclosed the real names of the two persons 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.

Analysis of Transparency Tools In our transparency tool analysis, we analyzed the data provided by 22 ad tech companies—a subset of all advertisement related companies. Still, we were able to identify different transparency approaches. The analyzed profiles do not include all data a company might collect (e.g., not all interest segments or all demographic information), and hence our classification might be false at some points. However, omitting the four cases in which we were unsure would not fundamentally change the overall results and findings.

In our user study, we surveyed participants from the US only. We decided *not* to recruit EU residents since we would

have had to provide the online questionnaire in various languages to avoid any bias because users do not take the questionnaire in their native language. Furthermore, we expect that US residents have similar needs when it comes to transparency in online advertisements. Even if they do not have the tools granted by the GDPR—as stated by P-53: “[Ads] are terrifying. We need GDPR in the US, at least as a first step”. After all, the online tools we tested are all available to any user. Our company survey is based on a small subset of companies willing to participate in our study and telephone interviews. Therefore, the views they presented might not be fully representative of the industry as a whole. Still, we identified a diverse set of opinions and hope that future work can broaden the empirical basis of our results.

4.3.2 Conclusion

In this chapter, we provided an overview of different approaches to how ad tech companies implemented their subject access request process, how users respond to data provided by these requests, and highlighted some challenges companies face when designing the processes.

Subject Access Request Process 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 companies grant access to personal data highlight 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.

Assessment of Transparency tools The ad industry tries to provide more transparency about its practices and the data collected in different ways. We studied implementations of new transparency and data access possibilities in the online advertising industry. By analyzing different transparency approaches of ad-tech companies, we identified three conceptual types of data companies provide to users, if they ask for access: (1) tracking data, (2) segment data, and (3) raw technical data. Our research shows that not all companies disclose the necessary information and that many do it in a way that is not user-friendly. The participants in our user study struggled to understand and

interpret the personal data they received after they had asked for access, especially if confronted with low-level technical data. Most users rated the provided data to be helpful ($> 50\%$), while “segment data” was the most popular category. Furthermore, we found that a large proportion of users (65%) do not trust that companies provide all collected data upon request. When it comes to the identification of companies that provide a standard ad banner, we found that only 24% of users would correctly identify the ad network, while 46% named the advertised products company as an ad provider. We surveyed data protection officers in different companies active in the advertisement ecosystem to understand their perspectives better. Participants reported that there were technical hurdles rooted in the complexity of the ecosystem that make it hard to disclose exact information. Most companies in the interviews and almost half in the survey (42%) stated that they receive less SARs than expected and 63% of participating companies expressed their wish for more guidance when designing SAR processes. We also found that companies still primarily focus on compliance instead of transparency for users. Regulatory authorities and industry associations, therefore, need to develop clear guidelines and consistent consumer-facing portals to improve the situation.

CHAPTER 5

CONCLUSION

In this chapter, we critically discuss the findings of this thesis. After a short discussion of possible implications of this thesis (Section 5.1), we briefly highlight future work that can build upon this thesis (Section 5.2) and provide an overall conclusion of our work (Section 5.3).

5.1 Possible Impact and Discussion

Modern online advertisement is complicated and not easy to grasp. Most websites confront their users with ads, yet

users do not know about the impact of these ads on their privacy, nor should they have to. The inner workings of the ecosystem raise several ethical questions. The most serious one is certainly if the tracking of millions of users is worth the slightly increased click rates if the basis of an ad is a behavioral profile. Often without any consent, ad tech companies intrude users' privacy for their gain.

In this work, we have shown that, from the user's point of view, current transparency mechanisms are not working. This lack of transparency might further fuel the distrust of users towards ad tech companies. A lack of trust and transparency might result in broad negative views of ads, in contrast to the stated neutral view in our experiment, which ultimately could hurt the industry. A trend towards paid services—that is free of ads and tracking—can be observed, which is an alternative way for service providers to generate revenue. Examples range from news websites [182, 183] over streaming services [70, 170], to mobile applications. Furthermore, users seem to distance themselves from services that came under public scrutiny for privacy violations (e.g., users used *Facebook* less after the *Cambridge Analytica* scandal [77]). However, as many services still rely on the revenue of ads and the fact that the strong lobby of online advertisers still fights to loosen GDPR requirements, a tipping point cannot be expected soon.

This thesis highlights open challenges that the GDPR did not solve or that need to be clarified. From a governance perspective, this thesis provides two main takeaways. First, the current inaccuracies regarding the GDPR right to access and right to data portability lead to a situation that is unsatisfactory for users and (ad tech) companies alike. Companies need guidelines that help them to determine what kind of data they should provide upon request. Furthermore, they need help to resolve the issue of checking the identity of users if they only have a pseudonymous tracking ID for each user. Secondly, the legal uncertainty in the usage of fourth and further parties should be addressed. Services in the modern Web often depend on third-party code. As shown in this thesis, this third-party code can load several partners, non deterministically. Even if aware of this practice, the service providers cannot control it, and aside from raising awareness, there should be legal certainty who is liable in cases if any of these consequentially loaded parties violates current legislation.

5.2 Future Work

In this thesis, we shed light on the implications of the GDPR on the online advertisement ecosystem and beyond. While much work went into understanding the impact of

the GDPR on the Web, there are still several unanswered questions in that regard. Furthermore, other domains have not yet been sufficiently explored. Some examples of potential future work in this area include:

1. Did mobile application changed their practices after enforcement of the GDPR (e.g., are less tracking libraries used)?,
2. What are the requirements to concepts that allow website providers to ensure that their services only include GDPR adequate parties (e.g., by loading dependencies) and how they can audit this?, or
3. What is the impact of different measurement setups on phenomena of interest, and what are the particular limitations of each approach?

The latter is an important question to consider—especially given the current reproducibility / comparability crisis in the web measurement community (see also Section 3.1). To the best of our knowledge, there is no direct comparison of the impact of different measurement approaches (e.g., full browser vs. headless browser vs. firing a simple get request from Python). Without a reliable comparison of different approaches, there is virtually no (scientific) argument to use one or the other. Nowadays, such decisions are often made either to decrease overhead (i.e., to increase measurement scaling) or by picking a common framework (e.g., *OpenWPM*) accepted in the community. Such studies

could sustainably effect future measurement studies and aid the (privacy) measurement community towards more comparable measurement results. Especially our results on vertical scaling of measurement approaches have shown that simple design choices can have a huge impact on the results.

Our analysis of the right to access created several open questions and challenges that should be addressed by research and normative institutions alike. Some of them are:

1. How can companies unmistakably identify users that perform SARs if the company only has a pseudonymous identifier?,
2. Which concepts should (small and medium-sized) companies use if they want to design or update their SAR processes? Can universal guidelines be developed?, or

From our point of view, the community lacks a real understanding of the transparency needs of users. For example, there is no academic knowledge which information privacy concerned users want/expect to see if they visit a website. Based on this lack of knowledge, several further questions arise that have not been answered yet. For example:

1. Are cookie-consent banners a suitable solution to inform users?

2. To which extent want users to control to block/limit online tracking and targeted ads? Is the current “block all or none” approach still timely and in line with users’ needs (e.g., by installing an adblocker)?
3. Is there an intuitive way to explain to users how online advertisers use their data and how can users gain actual control about these data flows?

5.3 Thesis Conclusion

The GDPR was the first serious attempt to protect users’ privacy by regulating data collection and usage of personal data. While the new legislation limits the business practices of companies, it provides several new rights to users. Online advertisement is one of the ecosystems that is particularly impacted by the GDPR as it heavily relies on data collection and processing (e.g., to provide targeted ads). Furthermore, the ecosystem is composed of several entities (e.g., ad providers, data brokers, companies that fight ad fraud, or ad auctioneers), which all have to share data, often without the user’s knowledge. Therefore, an analysis of the impact of the GDPR on this ecosystem yields several challenging and exciting research questions. This thesis provides answers to some of them.

First, we measured the technical impact that the GDPR had on essential parts of the online advertisement ecosystem. We have shown that the immediate effects of the new legislation are not measurable in all aspects of the ecosystem. On the one hand, this highlights the challenges of measuring effects of complex legislation (i.e., the GDPR) in a similarly complex environment (i.e., the Web) and, on the other hand, that not all expectations of privacy advocates towards the new legislation became a reality (e.g., tracking did not notably reduce). Our results show that overall fewer companies participate in the sharing economy of personal data but that the market-dominating players (e.g., *Google* or *Amazon*) even gained market shares due to this development. Furthermore, our experiments have shown that commonly used measurement setups might only present a limited view on a phenomenon of interest and that they should be revised. We have demonstrated that a vertical measurement setup, in contrast to a horizontal one, shows more stable results for a website's use of privacy-invasive technologies and also finds more of them.

Our human-centric experiments, as well as the surveyed literature, have revealed usability problems with the new rights to access and data portability. Companies, especially small and medium-sized ones, struggle to design processes that are compliant with the new legislation. As a result, these processes often tend to be complicated and time-

consuming for both users and companies. Furthermore, the success rate—in terms of companies that grant access to the collected data—was very low for the surveyed companies. As solving this is a challenge itself, we found that even if users get access to their data, they struggle to understand the presented data. Furthermore, users lack an understanding of the inner workings of the ad ecosystem to identify parties that get access to their data, which users should not need in the first place. In summary, the status quo of the implementations of the right to access and portability offer to (non-technical savvy) users only little benefit in regards to transparency of data usage. While some might argue that ad tech companies intend this, our results show that they are also unsure which data they should present users and how. This problem roots in the fact that there are no clear guidelines by industry near alliances, nor is there a clear consensus on how to interpret the new law in that regard.

BIBLIOGRAPHY

- [1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *Proceedings of the 21st ACM Conference on Computer and Communications Security, CCS '14*, pages 674–689, New York, NY, USA, 2014. ACM Press. doi:10.1145/2660267.2660347.
- [2] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens, and Bart Preneel. FPDetective: Dusting the Web for Finger-

- printers. In *Proceedings of the 20th ACM Conference on Computer and Communications Security, CCS '13*, pages 1129–1140, New York, NY, USA, 2013. ACM Press. doi:10.1145/2508859.2516674.
- [3] Adobe Inc. Flash & The Future of Interactive Content, 2017. Accessed: 2020-03-16. URL: <https://theblog.adobe.com/adobe-flash-update/>.
- [4] Simone Agostinelli, Fabrizio Maria Maggi, Andrea Marrella, and Francesco Sapio. Achieving GDPR Compliance of BPMN Process Models. In *Proceedings of the 31st Information Systems Engineering in Responsible Information Systems, CAiSE '19*, pages 10–22, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-21297-1_2.
- [5] Inc. Alexa Internet. Top Sites for Countries, 2018. Accessed: 2019-02-05. URL: <https://www.alexa.com/topsites/countries>.
- [6] Flora Amato, Luigi Coppolino, Salvatore D'Antonio, Nicola Mazzocca, Francesco Moscato, and Luigi Sgaglione. An Abstract Reasoning Architecture for Privacy Policies Monitoring. *Future Generation Computer Systems*, 106(1):393–400,

2020. doi:<https://doi.org/10.1016/j.future.2020.01.019>.
- [7] Amazon, Inc. Amazon’s Mechanical Turk, 2018. Accessed: 2020-03-16. URL: <https://www.mturk.com/>.
- [8] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook’s Explanations. In *Proceedings of the 24th Symposium on Network and Distributed System Security, NDSS ’18*, San Diego, CA, 2018. Internet Society. doi:10.14722/ndss.2018.23191.
- [9] Thibaud Antignac, Riccardo Scandariato, and Gerardo Schneider. Privacy Compliance Via Model Transformations. In *Proceedings of the 4th International Workshop on Privacy Engineering, IWPE ’18*, pages 120–126, Washington, DC, USA, April 2018. IEEE Computer Society. doi:10.1109/EuroSPW.2018.00024.
- [10] Apple Inc. Intelligent Tracking Prevention 2.3, 2019. Accessed: 2020-03-16. URL: <https://webkit.org/>

`blog/9521/intelligent-tracking-prevention-2-3/`.

- [11] Emma Arfelt, David Basin, and Søren Debois. Monitoring the GDPR. In Kazue Sako, Steve Schneider, and Peter Y. A. Ryan, editors, *Proceedings of the 24th European Symposium on Research in Computer Security*, ESORICS '19, pages 681–699, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-29959-0_33.
- [12] Guy Aridor, Yeon-Koo Che, William Nelson, and Tobias Salz. The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR. Technical report, Columbia University, 2020.
- [13] Article 29 Data Protection Working Party. Guidelines on the right to data portability. Technical Report 16 /EN WP 242, European Commission, December 2016.
- [14] Jef Ausloos and Pierre Dewitte. Shattering one-way mirrors—data subject access rights in practice. *International Data Privacy Law*, 8(1):4–28, 03 2018. doi:10.1093/idpl/ipy001.

- [15] Mika Ayenson, Dietrich James Wambach, Ashkan Soltani, Nathan Good, and Chris Jay Hoofnagle. Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning. *SSRN Electronic Journal*, 1(1), 2011. doi:10.2139/ssrn.1898390.
- [16] Paul Barford, Igor Canadi, Darja Krushevskaja, Qiang Ma, and S. Muthukrishnan. Adscape: Harvesting and Analyzing Online Display Ads. In *Proceedings of the 23rd International Conference on World Wide Web, WWW '14*, pages 597–608, Republic and Canton of Geneva, CHE, 2014. International World Wide Web Conferences Steering Committee. doi:10.1145/2566486.2567992.
- [17] Adam Barth. HTTP State Management Mechanism. RFC 6265, RFC Editor, 2011. doi:10.17487/RFC6265.
- [18] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. Tracing Information Flows Between Ad Exchanges Using Retargeted Ads. In *Proceedings of the 25th USENIX Security Symposium, USENIX Sec'16*, pages 481–496, Berkeley , CA, USA, 2016. USENIX Association.

- [19] Muhammad Ahmad Bashir, Umar Farooq, Maryam Shahid, Muhammad Fareed Zaffar, and Christo Wilson. Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers. In *Proceedings of the 25th Symposium on Network and Distributed System Security, NDSS'19*, San Diego, CA, USA, 2019. Internet Society. doi:10.14722/ndss.2019.23392.
- [20] Muhammad Ahmad Bashir and Christo Wilson. Diffusion of User Tracking Data in the Online Advertising Ecosystem. *Proceedings on Privacy Enhancing Technologies*, 4:85–103, 2018. doi:10.1515/popets-2018-0033.
- [21] David Basin, Søren Debois, and Thomas Hildebrandt. On Purpose and by Necessity: Compliance Under the GDPR. In *Proceedings of the 22nd Financial Cryptography and Data Security, FC '18*, pages 20–37, Cham, 2018. Springer International Publishing. doi:10.1007/978-3-662-58387-6_2.
- [22] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an Open Source Software for Exploring and Manipulating Networks. In *Proceedings of the 2nd International AAAI Conference on Weblogs and Social Media, ICWSM '09*, pages

- 361–362, San Jose, United States, 2009. AAAI. doi:10.13140/2.1.1341.1520.
- [23] Bits Of Freedom. My Data Done Right, 2019. Accessed: 2020-03-16. URL: <https://www.mydatadoneright.eu/>.
- [24] Coline Boniface, Imane Fouad, Nataliia Bielova, Cédric Lauradoux, and Cristiana Santos. Security Analysis of Subject Access Request Procedures. In *Proceedings of the 7th Annual Privacy Forum, APF '19*, pages 182–209, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-21752-5_12.
- [25] Brave Software Inc. Update on GDPR Complaint (RTB Ad Auctions), 2019. Accessed: 2020-03-16. URL: <https://www.brave.com/blog/update-rtb-ad-auction-gdpr/>.
- [26] Martin Brodin. A Framework for GDPR Compliance for Small- and Medium-Sized Enterprises. *European Journal for Security Research*, 4(2):243–264, 2019. doi:10.1007/s41125-019-00042-z.
- [27] Randolph E. Bucklin and Catarina Sismeiro. A Model of Web Site Browsing Behavior Estimated

- on Clickstream Data. *Journal of Marketing Research*, 40(3):249–267, 2003. doi:10.1509/jmkr.40.3.249.19241.
- [28] Matteo Cagnazzo, Thorsten Holz, and Norbert Pohlmann. GDPiRated—Stealing Personal Information On- and Offline. In *Proceedings of the 24th European Symposium on Research in Computer Security*, ESORICS '19, pages 367–386, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-29962-0_18.
- [29] California State Legislature. California Consumer Privacy Act of 2018, January 2018. URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.
- [30] Claude Castelluccia, Mohamed-Ali Kaafar, and Minh-Dung Tran. Betrayed by Your Ads!: Reconstructing User Profiles from Targeted Ads. *Proceedings on Privacy Enhancing Technologies*, pages 1–17, 2012. doi:10.1007/978-3-642-31680-7.
- [31] Sonia Chiasson, Yomna Abdelaziz, and Farah Chanchary. Privacy Concerns Amidst OBA and the Need for Alternative Models. *IEEE*

- Internet Computing*, 22(2):52–61, 2018. doi: 10.1109/MIC.2017.3301625.
- [32] Cliqz GmbH. Study: Google is the Biggest Beneficiary of the GDPR, 2018. Accessed: 2020-03-16. URL: <https://cliqz.com/en/magazine/study-google-is-the-biggest-beneficiary-of-the-gdpr>.
- [33] Cliqz GmbH. WhoTracks.me Data—Tracker database, 2018. Accessed: 2020-03-16. URL: <https://github.com/cliqz-oss/whotracks.me/tree/master/whotracksme/data>.
- [34] Commission Nationale de l’Informatique et des Libertés. Transmission des données à des partenaires à des fin de prospection électronique : quels sont les principes à respecter?, 2018. Accessed: 2020-03-16. URL: <https://www.cnil.fr/fr/transmission-des-donnees-des-partenaires-des-fin-de-prospection-electronique-quels-sont-les>.
- [35] Commission Nationale de l’Informatique et des Libertés. Deliberation of the Restricted Committee SAN-2019-001 of 21 January 2019 pronouncing a financial sanction against GOOGLE LL, 2019. Accessed: 2020-03-16. URL: <https://www.cnil.fr/>

sites/default/files/atoms/files/san-2019-001.pdf.

- [36] CONSENT project. Consent report summary, 2017. Accessed: 2020-03-16. URL: https://cordis.europa.eu/result/rcn/140471_en.html.
- [37] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. Measuring Cookies and Web Privacy in a Post-GDPR World. In *Proceedings of the 20th Conference on Passive and Active Measurement, PAM '19*, pages 258–270, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-15986-3_17.
- [38] Data Protection Working Party. Opinion 2/2010 on Online Behavioural Advertising, 2010. Accessed: 2020-03-16. URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp171_en.pdf.
- [39] Data Transfer Project. Data Transfer Project, 2018. Accessed: 2020-03-16. URL: <https://datatransferproject.dev/>.
- [40] Datastreams.io. Datastreams Platform, 2020. Accessed: 2020-03-16. URL: <https://www.datastreams.io/>.

- [41] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proceedings of the 25th Symposium on Network and Distributed System Security, NDSS '19*, San Diego, CA, USA, 2019. Internet Society. doi:10.14722/ndss.2019.23378.
- [42] Mariano Di Martino, Pieter Robyns, Winnie Weyts, Peter Quax, Wim Lamotte Lamotte, and Ken Andries. Personal Information Leakage by Abusing the GDPR "Right of Access". In *Proceedings of the 15th Symposium on Usable Privacy and Security, SOUPS '19*, pages 371—386, Berkeley, CA, USA, 2019. USENIX Association. doi:10.5555/3361476.3361504.
- [43] Digital Advertising Alliance. DAA Self-Regulatory Principles, 2018. Accessed: 2020-03-16. URL: <https://digitaladvertisingalliance.org/principles>.
- [44] Digital Advertising Alliance. Your Ad Choices, 2018. Accessed: 2020-03-16. URL: <https://optout.aboutads.info/?c=2&lang=EN>.

- [45] Dilecy GmbH. Dilecy | Your data, 2020. Accessed: 2020-03-16. URL: <https://dilecy.eu/>.
- [46] Claire Dolin, Ben Weinshel, Shawn Shan, Chang Min Hahn, Euirim Choi, Michelle L. Mazurek, and Blase Ur. Unpacking Perceptions of Data-Driven Inferences Underlying Online Targeting and Personalization. In *Proceedings of the 36th ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '18*, pages 1–12, New York, NY, USA, 2018. ACM Press. doi:10.1145/3173574.3174067.
- [47] EasyList. Easyprivacy, 2019. Accessed: 2020-03-16. URL: <https://easylist.to/easylist/easyprivacy.txt>.
- [48] eMarketer. Ad Blocking in the US: eMarketer’s Updated Estimates and Forecast for 2014–2018, 2017. Accessed: 2020-03-16. URL: <https://www.emarketer.com/Report/Ad-Blocking-US-eMarketers-Updated-Estimates-Forecast-20142018/2002044>.
- [49] Steven Englehardt and Arvind Narayanan. Online Tracking: A 1-Million-Site Measurement and Analysis. In *Proceedings of the 23rd ACM Conference on*

Computer and Communications Security, CCS '16, pages 1388–1401, New York, NY, USA, 2016. ACM Press. doi:10.1145/2976749.2978313.

- [50] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W. Felten. Cookies That Give You Away: The Surveillance Implications of Web Tracking. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, pages 289–299, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. doi:10.1145/2736277.2741679.
- [51] José Estrada-Jiménez, Javier Parra-Arnau, Ana Rodríguez-Hoyos, and Jordi Forné. Online Advertising: Analysis of Privacy Threats and Protection Approaches. *Computer Communications*, 100:32–51, 2017. doi:10.1016/j.comcom.2016.12.016.
- [52] European Commission. Adequacy Decisions—How the EU Determines If a Non-EU Country has an Adequate Level of Data Protection, 2020. Accessed: 2020-03-16. URL: <https://ec.europa.eu/info/law/law-topic/data->

protection/international-dimension-data-protection/adequacy-decisions_en.

- [53] European Court of Justice. Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein vs Wirtschaftsakademie Schleswig-Holstein GmbH, - Case C-210/16, 2018. URL: <http://curia.europa.eu/juris/document/document.jsf?text=&docid=202543&doclang=EN&d&part=1&cid=341550>.
- [54] European Interactive Digital Advertising Alliance. Your Online Choices, 2018. Accessed: 2020-03-16. URL: <http://www.youronlinechoices.com>.
- [55] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. Large-scale Readability Analysis of Privacy Policies. In *Proceedings of the 16th ACM International Conference on Web Intelligence, WI '17*, pages 18–25, New York, NY, USA, 2017. ACM Press. doi:10.1145/3106426.3106427.
- [56] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. Tracking Personal Identifiers Across the Web. In *Proceedings of the 17th Conference on Passive and Active Measurement, PAM '16*, pages 30–41, Cham, 2016. Springer In-

- ternational Publishing. doi:10.1007/978-3-319-30505-9_3.
- [57] Pietro Ferrara, Luca Olivieri, and Fausto Spoto. Tailoring Taint Analysis to GDPR. In *Proceedings of the 6th Annual Privacy Forum, APF '18*, pages 63–76, Cham, 2018. Springer International Publishing. doi:10.1007/978-3-030-02547-2_4.
- [58] Fingerprint.js. The Most Advanced Open-Source Fraud Detection JS Library, 2019. Accessed: 2020-03-16. URL: <https://fingerprintjs.com/>.
- [59] Uwe Flick. *The SAGE Handbook of Qualitative Data Analysis*. Sage Publications Ltd., Thousand Oaks, CA, USA, 2014.
- [60] David Formby, Preethi Srinivasan, Andrew Leonard, Jonathan Rogers, and Raheem Beyah. Who’s in Control of Your Control System? Device Fingerprinting for Cyber-Physical Systems. In *Proceedings of the 22nd Symposium on Network and Distributed System Security, NDSS '16*, San Diego, California, USA, 2016. Internet Society. doi:10.14722/ndss.2016.23142.
- [61] Imane Fouad, Nataliia Bielova, Arnaud Legout, and Natasa Sarafijanovic-Djukic. Missed by Fil-

- ter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. *Proceedings on Privacy Enhancing Technologies*, 2:499—518, July 2020. doi:10.2478/popets-2020-0038.
- [62] Gertjan Franken, Tom Van Goethem, and Wouter Joosen. Who Left Open the Cookie Jar? A Comprehensive Evaluation of Third-party Cookie Policies. In *Proceedings of the 27th USENIX Security Symposium*, USENIX Sec '18, pages 151–168, Berkeley, CA, USA, 2018. USENIX Association.
- [63] FutureScot. An unexpected benefit of gdpr; it makes the web much faster, 2018. Accessed: 2020-03-16. URL: <http://futurescot.com/an-unexpected-benefit-of-gdpr-it-makes-the-web-much-faster/>.
- [64] GDPR Implementation Working Group. Data Subject Requests. Technical Report Working Paper 04/2018 v1.0, IAB Europe, April 2018. URL: https://www.iabeurope.eu/wp-content/uploads/2018/04/20180406-IABEU-GIG-Working-Paper04_Data-Subject-Requests.pdf.
- [65] GEuropean Court of Justice. Case C-362/14—Maximillian Schrems vs. Data Protection Com-

- missioner, 2020. Accessed: 2020-07-20. URL: <http://curia.europa.eu/juris/document/document.jsf?text=&docid=169195&doclang=EN>.
- [66] Global Stats. Screen Resolution Stats, 2019. Accessed: 2020-03-16. URL: <http://gs.statcounter.com/screen-resolution-stats>.
- [67] Cliqz GmbH. GDPR—What happened?, 2018. Accessed: 2020-03-16. URL: <https://whotracks.me/blog/gdpr-what-happened.html>.
- [68] Alejandro Gómez-Boix, Pierre Laperdrix, and Benoit Baudry. Hiding in the Crowd: An Analysis of the Effectiveness of Browser Fingerprinting at Large Scale. In *Proceedings of the 27th International Conference on World Wide Web, WWW '18*, pages 309—318, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. doi:10.1145/3178876.3186097.
- [69] Roberto Gonzalez, Lili Jiang, Mohamed Ahmed, Miriam Marciel, Ruben Cuevas, Hassan Metwalley, and Saverio Niccolini. The cookie recipe: Untangling the Use of Cookies in the Wild. In *Proceedings of the 1st Network Traffic Measurement and*

Analysis Conference, TMA '17, pages 1–9, Piscataway, NJ, USA, 2017. IEEE Computer Society. doi:10.23919/TMA.2017.8002896.

- [70] Google Inc. YouTube Premium, 2020. Accessed: 2020-03-16. URL: <https://www.youtube.com/premium>.
- [71] Google LLC. Turn “Do Not Track” On or Off, 2019. Accessed: 2020-03-16. URL: <https://support.google.com/chrome/answer/2790761>.
- [72] Government Digital Service. Countries in the EU and EEA, 2019. Accessed: 2020-03-16. URL: <https://www.gov.uk/eu-eea>.
- [73] Inge Graef, Martin Husovec, and Nadezhda Purtova. Data Portability and Data Control: Lessons for an Emerging Concept in EU Law. *German Law Journal*, 19(6):1359—1398, 2018. doi:10.1017/S2071832200023075.
- [74] Samuel Grogan and Aleecia M. McDonald. Access Denied! Contrasting Data Access in the United States and Ireland. *Proceedings of the Proceedings on Privacy Enhancing Technologies*, 3(23):191–211, 07 2016. doi:10.1515/popets-2016-0023.

- [75] Aric Hagberg, Daniel Schult, and Pieter Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, SciPy '08, pages 11 – 15, Pasadena, CA USA, 2008.
- [76] Daniel Hausknecht, Jonas Magazinius, and Andrei Sabelfeld. May I?-Content Security Policy Endorsement for Browser Extensions. In *Proceedings of the 12th Conference on Detection of Intrusions and Malware & Vulnerability Assessment*, DIMVA '15, pages 261–281, Cham, 2015. Springer International Publishing. doi:10.1007/978-3-319-20550-2_14.
- [77] Alex Hern. Facebook usage falling after privacy scandals, data suggests , 2019. Accessed: 2020-03-16. URL: <https://www.theguardian.com/technology/2019/jun/20/facebook-usage-collapsed-since-scandal-data-shows>.
- [78] Alex Hern. Google Fined Record £44m by French Data Protection Watchdog, January 2019. Accessed: 2020-03-16. URL: <https://www.theguardian.com/technology/2019/jan/21/google-fined-record-44m-by-french-data-protection-watchdog>.

- [79] Paul De Hert, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, and Ignacio Sanchez. The Right to Data Portability in the GDPR: Towards User-Centric Interoperability of Digital Services. *Computer Law & Security Review*, 34(2):193–203, 2018. doi:10.1016/j.clsr.2017.10.003.
- [80] Higher Regional Court, Düsseldorf, Germany. Opinion of Advocate General Bobek on Fashion ID GmbH & Co. KG vs Verbraucherzentrale NRW eV—Case C-40/1, December 2018. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=ecli:ECLI:EU:C:2018:1039>.
- [81] Mireille Hildebrandt. The Dawn of a Critical Transparency Right for the Profiling Era. *Digital Enlightenment Yearbook*, pages 41–56, 2012. doi:10.3233/978-1-61499-057-4-41.
- [82] Kalle Hjerpe, Jukka Ruohonen, and Ville Lepänen. The General Data Protection Regulation: Requirements, Architectures, and Constraints. In *Proceedings of the 27th IEEE International Requirements Engineering Conference*, RE '19, pages 265–275, Washington, DC, USA, 2019. IEEE Computer Society. doi:10.1109/RE.2019.00036.

- [83] Falk Howard. Computer Intrusions and Attacks. *The Electronic Library*, 17(2):115–119, Jan 1999. doi:10.1108/02640479910329635.
- [84] Dominik Huth, Laura Stojko, and Florian Matthes. A Service Definition for Data Portability. In *Proceedings of the 21st International Conference on Enterprise Information Systems, ICEIS'19*, pages 169–176, Setúbal, Portugal, 2019. SciTePress. doi:10.5220/0007677101690176.
- [85] IAB Europe. European digital advertising market has doubled in size in 5 years, 2017. Accessed: 2020-03-16. URL: <https://www.iabeurope.eu/research-thought-leadership/resources/iab-europe-report-adex-benchmark-2017-report/>.
- [86] IAB Europe. Iab europe transparency & consent framework policies, 2018. Accessed: 2020-03-16. URL: <http://www.iabeurope.eu/tcfdocuments/documents/legal/currenttcfpolicyFINAL.pdf>.
- [87] Muhammad Ikram, Rahat Masood, Gareth Tyson, Mohamed Ali Kaafar, Noha Loizon, and Roya En-safi. The Chain of Implicit Trust: An Analysis of the Web Third-Party Resources Loading. In *Proceedings of the 28th International Conference on*

World Wide Web, WWW '19, pages 2851—2857, Republic and Canton of Geneva, CHE, 2019. International World Wide Web Conferences Steering Committee. doi:10.1145/3308558.3313521.

- [88] Interactive Advertising Bureau. Internet Advertising Revenue Report, 2017. Accessed: 2020-03-16. URL: https://www.iab.com/wp-content/uploads/2018/05/IAB-2017-Full-Year-Internet-Advertising-Revenue-Report.REV2_.pdf.
- [89] International Chamber of Commerce UK. *Cookie Guide*. International Chamber of Commerce UK, 2012. URL: https://www.cookielaw.org/wp-content/uploads/2019/12/icc_uk_cookiesguide_revnov.pdf.
- [90] Internet World Stats. Top 20 Countries with the Highest Number of Internet Users, 2018. Accessed: 2019-02-05. URL: <https://www.internetworldstats.com/top20.htm>.
- [91] Luca Invernizzi, Kurt Thomas, Alexandros Kapravelos, Oxana Comanescu, Jean-Michel Picod, and Elie Bursztein. Cloak of Visibility: Detecting When Machines Browse a Different Web. In *Proceedings of the 37th IEEE Symposium on Secu-*

- ity and Privacy*, S&P '16, pages 743–758, Piscataway, NJ, USA, 2016. IEEE Computer Society. doi:10.1109/SP.2016.50.
- [92] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. Tracing Cross Border Web Tracking. In *Proceedings of the 18th ACM SIGCOMM Internet Measurement Conference, IMC '18*, pages 329–342, New York, NY, USA, 2018. ACM Press. doi:10.1145/3278532.3278561.
- [93] IPLocation. Where is Geolocation of an IP Address?, 2019. Accessed: 2020-03-16. URL: <https://www.iplocation.net/>.
- [94] Tim Jackson. The Bug in your PC is a smart cookie. *Financial Times*, 1996. Accessed: 2019-11-11. URL: [https://archive.org/stream/FinancialTimes1996UKEnglish/Feb%2012%201996%2C%20Financial%20Times%2C%20%2312%2C%20UK%20\(en\)#page/n29/mode/2up](https://archive.org/stream/FinancialTimes1996UKEnglish/Feb%2012%201996%2C%20Financial%20Times%2C%20%2312%2C%20UK%20(en)#page/n29/mode/2up).
- [95] Musa J Jafar and Amjad Abdullat. Exploratory Analysis of the Readability of Information Privacy Statement of the Primary Social Networks. *Journal of Business & Economics Research*, 7(12):123–142, 2009. doi:10.19030/jber.v7i12.2371.

- [96] Carlos Jensen and Colin Potts. Privacy Policies As Decision-making Tools: An Evaluation of Online Privacy Notices. In *Proceedings of the 22th ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '04*, pages 471–478, New York, NY, USA, 2004. ACM Press. doi:10.1145/985692.985752.
- [97] Qiwei Jia, Lu Zhou, Huaxin Li, Ruoxu Yang, Suguo Du, and Haojin Zhu. Who Leaks My Privacy: Towards Automatic and Association Detection with GDPR Compliance. In *Proceedings of the 14th International Conference on Wireless Algorithms, Systems, and Applications, WASA '19*, pages 137–148, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-23597-0_11.
- [98] Garrett Johnson and Scott Shriver. Privacy & Market Concentration: Intended & Unintended Consequences of the GDPR. Technical report, Questrom School of Business, 2020. doi:10.2139/ssrn.3477686.
- [99] Hugo Jonker, Benjamin Krumnow, and Gabry Vlot. Fingerprint Surface-Based Detection of Web Bot Detectors. In *Proceedings of the 24th European Symposium on Research in Computer Security, ES-*

- ORICS '19, pages 586–605, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-29962-0_28.
- [100] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Joseph M. Pujol. WhoTracks.Me: Monitoring the Online Tracking Landscape at Scale. *CoRR*, abs/1804.08959, 2018. URL: <http://arxiv.org/abs/1804.08959>, arXiv:1804.08959.
- [101] Mohammad Taha Khan, Joe DeBlasio, Geoffrey M. Voelker, Alex C. Snoeren, Chris Kanich, and Narseo Vallina-Rodriguez. An Empirical Analysis of the Commercial VPN Ecosystem. In *Proceedings of the 18th ACM SIGCOMM Internet Measurement Conference, IMC '18*, pages 443–456, New York, USA, 2018. ACM Press. doi:10.1145/3278532.3278570.
- [102] I Luk Kim, Weihang Wang, Yonghwi Kwon, Yunhui Zheng, Yousra Aafer, Weijie Meng, and Xianguyu Zhang. AdBudgetKiller: Online Advertising Budget Draining Attack. In *Proceedings of the 27th International Conference on World Wide Web, WWW '18*, pages 297–307, Republic and Canton of Geneva, CHE, 2018. International

World Wide Web Conferences Steering Committee.
doi:10.1145/3178876.3186096.

- [103] Radhesh Krishnan Konoth, Emanuele Vineti, Veelasha Moonsamy, Martina Lindorfer, Christopher Kruegel, Herbert Bos, and Giovanni Vigna. MineSweeper: An In-depth Look into Drive-by Cryptocurrency Mining and Its Defense. In *Proceedings of the 25th ACM Conference on Computer and Communications Security, CCS '18*, pages 1714–1730, New York, NY, USA, 2018. ACM Press. doi:10.1145/3243734.3243858.
- [104] Martin Koop, Eri Tews, and Stefan Katzenbeisser. In-Depth Evaluation of Redirect Tracking and Link Usage. *Proceedings on Privacy Enhancing Technologies*, 4:394–413, July 2020. doi:10.2478/popets-2020-0079.
- [105] Korea Rep. Personal Information Protection Act, September 2011. URL: <http://www.pipc.go.kr/cmt/english/functions/infoCommunication.do>.
- [106] David M. Kristol. HTTP Cookies: Standards, Privacy, and Politics. *ACM Transactions on Internet Technology*, 1(2):151–198, November 2001. doi:10.1145/502152.502153.

- [107] Shubham Kumar, Siddharth Rautaray, and Manjusha Pandey. Malvertising: A Case Study Based on Analysis of Possible Solutions. In *Proceedings of the 1st International Conference on Inventive Computing and Informatics, ICICI '17*, pages 288–291, Piscataway, NJ, USA, 2017. IEEE Computer Society. doi:10.1109/ICICI.2017.8365356.
- [108] Andreas Kurtz, Hugo Gascon, Tobias Becker, Konrad Rieck, and Felix C. Freiling. Fingerprinting Mobile Devices Using Personalized Configurations. *Proceedings of the Proceedings on Privacy Enhancing Technologies*, 2016(1):4–19, 2016. doi:10.1515/popets-2015-0027.
- [109] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. Browser Fingerprinting: A Survey. *ACM Transactions on the Web*, 14(2), 2020. doi:10.1145/3386040.
- [110] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the 26th Symposium on Network and Distributed System Security, NDSS '19*, San

Diego, California, USA, 2019. Internet Society.
doi:10.14722/ndss.2019.23386.

- [111] Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. Why Johnny Can't Opt out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising. In *Proceedings of the 30th ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 589–598, New York, NY, USA, 2012. ACM. doi: 10.1145/2207676.2207759.

- [112] Timothy Libert, Lucas Graves, and Rasmus Kleis Nielsen. Changes in Third-Party Content on European News Websites after GDPR. Technical report, Reuters Institute for the Study of Journalism, Oxford, UK, 2018. Accessed: 2019-10-05. URL: 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.

- [113] Timothy Libert and Kleis Nielsen Nielsen. Third-Party Web Content on EU News Sites: Potential Challenges and Paths to Privacy. Technical report, Reuters Institute for the Study of Journalism, 2018.

- [114] LimeSurvey Project Hamburg. Limesurvey: An open source survey tool, 2012. Accessed: 2020-03-16. URL: <https://www.limesurvey.org/>.
- [115] Bin Liu, Anmol Sheth, Udi Weinsberg, Jaideep Chandrashekar, and Ramesh Govindan. AdReveal: Improving Transparency into Online Targeted Advertising. In *Proceedings of the 12th ACM SIGCOMM Workshop on Hot Topics in Networks, HotNets '12*, pages 1–7, New York, NY, USA, 2013. ACM Press. doi:10.1145/2535771.2535783.
- [116] Miguel Malheiros, Charlene Jennett, Sneha Patel, Sacha Brostoff, and Martina Angela Sasse. Too Close for Comfort: A Study of the Effectiveness and Acceptability of Rich-media Personalized Advertising. In *Proceedings of the 30th ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, pages 579–588, New York, NY, USA, 2012. ACM Press. doi:10.1145/2207676.2207758.
- [117] Célestin Matte, Nataliia Bielova, and Cristiana Santos. Do Cookie Banners Respect My Choice? Measuring Legal Compliance of Banners from IAB Europe’s Transparency and Consent Framework. In *Proceedings of the 41st IEEE Symposium on*

Security and Privacy, S&P '20, pages 1612–1630, Piscataway, NJ, USA, 2020. IEEE Computer Society. doi:10.1109/SP40000.2020.00076.

- [118] Zachary Matthews and Natalija Vlajic. Poster: Can Browser Add-Ons Protect Your Children from Online Tracking? In *Proceedings of the 25th ACM Conference on Computer and Communications Security, CCS '18*, pages 2243–2245, New York, NY, USA, 2018. ACM Press. doi:10.1145/3243734.3278498.
- [119] MaxMind Inc. GeoIP Databases & Services, 2019. Accessed: 2020-03-16. URL: <https://www.maxmind.com/en/geop2-services-and-databases>.
- [120] Jonathan R. Mayer and John C. Mitchell. Third-Party Web Tracking: Policy and Technology. In *Proceedings of the 33rd IEEE Symposium on Security and Privacy, S&P '12*, pages 413–427, Piscataway, NJ, USA, 2012. IEEE Computer Society. doi:10.1109/SP.2012.47.
- [121] McAfee LLC. Customer url ticketing system, 2019. Accessed: 2020-03-16. URL: <https://trustedsource.org/>.

- [122] Aleecia McDonald and Jon M. Peha. Track Gap: Policy Implications of User Expectations for the ‘Do Not Track’ Internet Privacy Feature. Technical report, Social Science Research Network, 2011.
- [123] Aleecia M. McDonald and Lorrie Faith Cranor. Americans’ Attitudes About Internet Behavioral Advertising Practices. In *Proceedings of the 9th ACM Workshop on Privacy in the Electronic Society*, WPES ’10, pages 63–72, New York, NY, USA, 2010. ACM Press. doi:10.1145/1866919.1866929.
- [124] William Melicher, Mahmood Sharif, Joshua Tan, Lujjo Bauer, Mihai Christodorescu, and Pedro Giovanni Leon. (Do Not) Track Me Sometimes: Users’ Contextual Preferences for Web Tracking. *Proceedings of the Proceedings on Privacy Enhancing Technologies*, 2016(2):135–154, 2016. doi:10.1515/popets-2016-0009.
- [125] Georg Merzdovnik, Markus Huber, Markus Buhov, Nick Nikiforakis, Sebastian Neuner, Martin Schmiedecker, and Edgar Weippl. Block Me If You Can: A Large-Scale Study of Tracker-Blocking Tools. In *Proceedings of the 2nd IEEE European Symposium on Security and Privacy*, EuroS&P ’17,

- pages 319–333, Piscataway, NJ, USA, 2017. IEEE Computer Society. doi:10.1109/EuroSP.2017.26.
- [126] Meik Michalke. koRpus: An R Package for Text Analysis, 2018. Version: 0.11-5. Accessed: 2020-03-16. URL: <https://reaktanz.de/?c=hacking&s=koRpus>.
- [127] MIT Technology Review. Campaigns to Track Voters with “Political Cookies”, 2019. Accessed: 2020-03-16. URL: <https://www.technologyreview.com/s/428347/campaigns-to-track-voters-with-political-cookies/>.
- [128] Dimitris Mitropoulos, Thodoris Sotiropoulos, Nikos Koutsovasilis, and Diomidis Spinellis. PDGuard: An Architecture for the Control and Secure Processing of Personal Data. *International Journal of Information Security*, 2019. doi:10.1007/s10207-019-00468-5.
- [129] Mozilla Corporation. Today’s Firefox Blocks Third-Party Tracking Cookies and Cryptomining by Default, 2019. Accessed: 2020-03-16. URL: <https://blog.mozilla.org/blog/2019/09/03/todays-firefox-blocks-third-party-tracking-cookies-and-cryptomining-by-default/>.

- [130] MyCustomer. Will GDPR Kill the Third-Party Data Market?, 2018. Accessed: 2020-03-16. URL: <https://www.mycustomer.com/marketing/data/will-gdpr-kill-the-third-party-data-market>.
- [131] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. PriVaricator: Deceiving Fingerprinters with Little White Lies. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, pages 820–830, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. doi:10.1145/2736277.2741090.
- [132] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting. In *Proceedings of the 34th IEEE Symposium on Security and Privacy, S&P '13*, pages 541–555, Piscataway, NJ, USA, 2013. IEEE Computer Society. doi:10.1109/SP.2013.43.
- [133] Clive Norris, Paul de Hert, Xavier L'Hoiry, and Antonella Galetta, editors. *The Unaccountable State of Surveillance*. Springer International Publishing, Cham, 2017. doi:10.1007/978-3-319-47573-8.

- [134] Clive Norris and Xavier L’Hoiry. *Exercising Citizen Rights Under Surveillance Regimes in Europe—Meta-analysis of a Ten Country Study*, pages 405–455. Springer International Publishing, Cham, 2017. doi:10.1007/978-3-319-47573-8_14.
- [135] Gabriel Nunes. Adobe is Helping Some 60 Companies Track People Across Devices, 2018. Accessed: 2020-03-16. URL: <https://www.neowin.net/news/adobe-is-helping-some-60-companies-track-people-across-devices>.
- [136] Lukasz Olejnik and Claude Castelluccia. To Bid or Not to Bid? Measuring the Value of Privacy in RTB. Technical report, Institut National De Recherche En Informatique Et en Automatique, Grenoble, 2014.
- [137] OneTrust LLC. Cookiepedia, 2019. Accessed: 2020-03-16. URL: <https://cookiepedia.co.uk/>.
- [138] OpenGDPR. OpenDSR, 2020. Accessed: 2020-03-16. URL: <https://github.com/opengdpr/OpenDSR>.
- [139] Mahieu René L. P., Hadi Asghari, and Michel van Eeten. Collectively Exercising the Right of Access: Individual Effort, Societal Effect. *Internet Policy Review*, 7(3), 2018. doi:10.14763/2018.3.927.

- [140] Xiang Pan, Yinzi Cao, and Yan Chen. I Do Not Know What You Visited Last Summer: Protecting Users from Third-party Web Tracking with TrackingFree Browser. In *Proceedings of the 21st Symposium on Network and Distributed System Security*, NDSS '15, San Diego, California, USA, 2015. Internet Society. doi:10.14722/ndss.2015.23163.
- [141] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. In *Proceedings of the 28th International Conference on World Wide Web*, WWW '19, pages 1432—1442, Republic and Canton of Geneva, CHE, 2019. International World Wide Web Conferences Steering Committee. doi:10.1145/3308558.3313542.
- [142] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. The Cost of Digital Advertisement: Comparing User and Advertiser Views. In *Proceedings of the 27th International Conference on World Wide Web*, WWW '18, pages 1479–1489, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. doi:10.1145/3178876.3186060.

- [143] Javier Parra-Arnau, Jagdish Prasad Achara, and Claude Castelluccia. MyAdChoices: Bringing Transparency and Control to Online Advertising. *ACM Transactions on the Web*, 11(1):7–47, 2017. doi:10.1145/2996466.
- [144] Personal Privacy Solutions. Tapmydata—Discover what personal data organisations hold about you, 2020. Accessed: 2020-03-16. URL: <https://tapmydata.com/>.
- [145] Angelisa C. Plane, Elissa M. Redmiles, Michelle L. Mazurek, and Michael Carl Tschantz. Exploring User Perceptions of Discrimination in Online Targeted Advertising. In *Proceedings of the 26th USENIX Security Symposium*, USENIX Sec '17, pages 935–951, Berkeley , CA, USA, 2017. USENIX Association.
- [146] Python Software Foundation. tldextract, 2019. Version: 2.2.1. Accessed: 2020-03-16. URL: <https://pypi.org/project/tldextract/>.
- [147] Ashwini Rao and Juergen Pfeffer. Data Siphoning Across Borders: The Role of Internet Tracking. In *Proceedings of the 1st IEEE International Conference on Trust, Privacy and Security in*

- Intelligent Systems and Applications*, TPS-ISA '19, pages 168–176, Washington, DC, USA, Dec 2019. IEEE Computer Society. doi:10.1109/TPS-ISA48467.2019.00028.
- [148] Ashwini Rao, Florian Schaub, and Norman M. Sadeh. What do they know about me? Contents and Concerns of Online Behavioral Profiles. *CoRR*, abs/1506.01675, 2015. URL: <http://arxiv.org/abs/1506.01675>, arXiv:1506.01675.
- [149] John W Ratcliff and David E Metzener. Pattern Matching: The Gestalt Approach. *Dr Dobbs Journal*, 13(7):46, 1988.
- [150] Joel Reardon, Álvaro Feal, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, and Serge Egelman. 50 Ways to Leak Your Data: An Exploration of Apps’ Circumvention of the Android Permissions System. In *Proceedings of the 28th USENIX Security Symposium*, USENIX Sec’16, pages 603–620, Santa Clara, CA, August 2019. USENIX Association.
- [151] Irwin Reyes, Primal Wijesekera, Joel Reardon, Amit Bar On Elazari, Abbas Razaghpanah, Narseo Vallina-Rodriguez, and Serge Egelman. “Won’t

- Somebody Think of the Children?” Examining COPPA Compliance at Scale. *Proceedings on Privacy Enhancing Technologies*, 2018(3):63–83, 2018. doi:10.1515/popets-2018-0021.
- [152] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and Defending Against Third-party Tracking on the Web. In *Proceedings of the 9th Conference on Networked Systems Design and Implementation*, NSDI '12, pages 155–168, Berkeley , CA, USA, 2012. USENIX Association. doi:10.5555/2228298.2228315.
- [153] Jan R uth, Torsten Zimmermann, Konrad Wolsing, and Oliver Hohlfeld. Digging into Browser-based Crypto Mining. In *Proceedings of the 18th ACM SIGCOMM Internet Measurement Conference*, IMC '18, pages 70–76, New York, NY, USA, 2018. ACM Press. doi:10.1145/3278532.3278539.
- [154] Johnny Ryan. The 3 Biggest Challenges in GDPR for Online Media & Advertising, 2017. Accessed: 2020-03-16. URL: <https://johnnyryan.wordpress.com/2017/07/19/gdpr-3-deep-challenges/>.

- [155] Takahito Sakamoto and Masahiro Matsunaga. After GDPR, Still Tracking or Not? Understanding Opt-Out States for Online Behavioral Advertising. In *Proceedings of the 5th International Workshop on Privacy Engineering, IWPE' 19*, pages 92–99, May 2019. doi:10.1109/SPW.2019.00027.
- [156] Sonam Samat, Alessandro Acquisti, and Linda Babcock. Raise the Curtains: The Effect of Awareness About Targeting on Consumer Attitudes and Purchase Intentions. In *Proceedings of the 13th Symposium on Usable Privacy and Security, SOUPS'17*, pages 299–319, Vancouver, BC, 2017. USENIX Association.
- [157] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. Can I Opt Out Yet?: GDPR and the Global Illusion of Cookie Control. In *Proceedings of the 14th ACM Asia Conference on Computer and Communications Security, AsiaCCS '19*, pages 340–351, New York, New York, USA, 2019. ACM Press. doi:10.1145/3321705.3329806.
- [158] Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lor-

- rie Faith Cranor. Watching Them Watching Me: Browser Extensions Impact on User Privacy Awareness and Concern. In *Proceedings of the 4th Workshop on Usable Security, USEC '16*, Reston, VA, 2016. Internet Society. doi: 10.14722/usec.2016.23017.
- [159] Klaus Schwab, Alan Marcus, JO Oyola, William Hoffman, and M Luzi. Personal data: The Emergence of a New Asset Class. Technical report, World Economic Forum, 2011.
- [160] Andrew D. Selbst and Julia Powles. Meaningful Information and the Right to Explanation. *International Data Privacy Law*, 7(4):233–242, November 2017. doi:10.1093/idpl/idx022.
- [161] Sharethrough Inc. Sharethrough User ID Validator, 2018. Accessed: 2020-03-16. URL: <https://integration.sharethrough.com/tools/user-validator>.
- [162] Muazzam Siddiqui, Morgan C. Wang, and Joochan Lee. Data Mining Methods for Malware Detection Using Instruction Sequences. In *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications, AIA '08*,

- pages 358–363, Anaheim, CA, USA, 2008. ACTA Press. doi:10.5555/1712759.1712825.
- [163] Jatinder Singh and Jennifer Cobbe. The Security Implications of Data Subject Rights. *IEEE Security & Privacy*, 17(6):21–30, 2019. doi:10.1109/MSEC.2019.2914614.
- [164] Sean Sirur, Jason R.C. Nurse, and Helena Webb. Are We There Yet? Understanding the Challenges Faced in Complying with the General Data Protection Regulation (GDPR). In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, MPS '18, pages 88–95, New York, NY, USA, 2018. ACM Press. doi:10.1145/3267357.3267368.
- [165] Rocky Slavin, Xiaoyin Wang, Mitra Bokaei Hosseini, James Hester, Ram Krishnan, Jaspreet Bhatia, Travis D. Breaux, and Jianwei Niu. Toward a Framework for Detecting Privacy Policy Violations in Android Application Code. In *Proceedings of the 38th International Conference on Software Engineering*, ICSE '16, pages 25–36, May 2016. doi:10.1145/2884781.2884855.

- [166] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. Clash of the Trackers: Measuring the Evolution of the Online Tracking Ecosystem. Technical report, Foundation for Research and Technology, Jul 2019. [arXiv:1907.12860](https://arxiv.org/abs/1907.12860).
- [167] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. Flash Cookies and Privacy. *SSRN Electronic Journal*, 1(1), 2009. [doi:10.2139/ssrn.1446862](https://doi.org/10.2139/ssrn.1446862).
- [168] Aditya K. Sood and Richard J. Enbody. Malvertising—Exploiting Web Advertising. *Computer Fraud & Security*, 2011(4):11–16, 2011. [doi:10.1016/S1361-3723\(11\)70041-0](https://doi.org/10.1016/S1361-3723(11)70041-0).
- [169] Jannick Kirk Sørensen and Sokol Kosta. Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. In *Proceedings of the 28th International Conference on World Wide Web, WWW '19*, pages 1590–1600, Republic and Canton of Geneva, CHE, 2019. International World Wide Web Conferences Steering Committee. [doi:10.1145/3308558.3313524](https://doi.org/10.1145/3308558.3313524).

- [170] Spotify AB. Spotify Premium, 2019. Accessed: 2020-03-16. URL: <https://www.spotify.com/en/premium/>.
- [171] Oleksii Starov and Nick Nikiforakis. Extended Tracking Powers: Measuring the Privacy Diffusion Enabled by Browser Extensions. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1481–1490, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. doi:10.1145/3038912.3052596.
- [172] Talend. The Majority of Businesses Surveyed are Failing to Comply with GDPR, 2018. Accessed: 2020-03-16. URL: <https://www.talend.com/about-us/press-releases/the-majority-of-businesses-are-failing-to-comply-with-gdpr-according-to-new-talend-research/>.
- [173] Tech Blog. Most Common User Agents, 2019. Accessed: 2020-03-16. URL: <https://techblog.willshouse.com/2012/01/03/most-common-user-agents/>.
- [174] Th Financial Times. How Top Health Websites are Sharing Sensitive Data with Advertisers,

November 2019. Accessed: 2019-12-05. URL: <https://www.ft.com/content/0fbf4d8e-022b-11ea-be59-e49b2a136b8d>.

- [175] The Drum. The Day After Tomorrow: When Ad-blockers and GDPR Kill All Adtech and Martech, 2017. Accessed: 2020-03-16. URL: <https://www.thedrum.com/opinion/2017/10/17/the-day-after-tomorrow-when-ad-blockers-and-gdpr-kill-all-adtech-and-martech>.
- [176] The European Parliament and the Council of the European Union. Directive 2009/136/ec of the european parliament and of the council of 25 november 2009 amending directive 2002/22/ec on universal service and users' rights relating to electronic communications networks and services, directive 2002/58/ec concerning the processing of personal data and the protection of privacy in the electronic communications sector and regulation (ec) no 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws.
- [177] The European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27

April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119/1, April 2016.

- [178] The Guardian. Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach, 2018. Accessed: 2020-03-16. URL: <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>.
- [179] The International Trade Administration. The Privacy Shield, 2019. Accessed: 2020-03-16. URL: <https://www.privacyshield.gov/>.
- [180] The New York Times. The Slow Death of ‘Do Not Track’, 2014. Accessed: 2020-03-16. URL: <http://www.nytimes.com/2014/12/27/opinion/the-slow-death-of-do-not-track.html>.
- [181] The New York Times. 5 Ways Facebook Shared Your Data, 2018. Accessed: 2020-03-16. URL: <https://www.nytimes.com/2018/12/19/technology/facebook-data-sharing.html>.

- [182] The New York Times. Digital and Home Delivery Subscriptions, 2020. Accessed: 2020-03-16. URL: <https://www.nytimes.com/subscription>.
- [183] The Washington Post. Subscribe to The Washington Post, 2020. Accessed: 2020-03-16. URL: <https://subscribe.washingtonpost.com/>.
- [184] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 4 Years of EU Cookie Law: Results and Lessons Learned. *Proceedings on Privacy Enhancing Technologies*, 2019(2):126–145, 2019. doi:10.2478/popets-2019-0023.
- [185] Triple Lift Inc. Your Individual Rights, 2018. Accessed: 2020-03-16. URL: <https://access.triplelift.com/>.
- [186] TRUSTe and Harris Interactive. Consumer Research Results—Privacy and Online Behavioral Advertising, 2011. Accessed: 2020-03-16. URL: <https://www.eff.org/files/truste-2011-consumer-behavioral-advertising-survey-results.pdf>.
- [187] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *Proceedings of the 8th Sym-*

- posium on Usable Privacy and Security*, SOUPS '12, pages 1–15, New York, NY, USA, 2012. ACM Press. doi:10.1145/2335356.2335362.
- [188] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. “Your Hashed IP Address: Ubuntu.”: Perspectives on Transparency Tools for Online Advertising. In *Proceedings of the 35th Annual Computer Security Applications Conference*, ACSAC '19, pages 702–717, New York, NY, USA, 2019. ACM Press. doi:10.1145/3359789.3359798.
- [189] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Beyond the Front Page: Measuring Third Party Dynamics in the Field. In *Proceedings of the 29th International Conference on World Wide Web*, WWW '20, Republic and Canton of Geneva, CHE, 2020. International World Wide Web Conferences Steering Committee. doi:10.1145/3366423.3380203.
- [190] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. A Study on Subject Data Access in Online Advertising after the GDPR. In *Proceedings of the 14th International Workshop on Data Privacy Management*, DPM '19,

pages 61–79, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-31500-9_5.

- [191] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Measuring the Impact of the GDPR on Data Sharing. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security, AsiaCCS '20*, New York, NY, USA, 2020. ACM Press. doi: 10.1145/3320269.3372194.

- [192] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. The Unwanted Sharing Economy: An Analysis of Cookie Syncing and User Transparency under GDPR. *ArXiv e-prints*, November 2018. arXiv: 1811.08660.

- [193] Tobias Urban, Dennis Tatang, Thorsten Holz, and Norbert Pohlmann. Towards Understanding Privacy Implications of Adware and Potentially Unwanted Programs. In *Proceedings of the 23rd European Symposium on Research in Computer Security, ESORICS '18*, pages 449–469, Cham, 2018. Springer International Publishing. doi:10.1007/978-3-319-99073-6_22.

- [194] Tobias Urban, Dennis Tatang, Thorsten Holz, and Norbert Pohlmann. Analyzing Leakage of Personal Information by Malware. *Journal of Computer Security*, 27(4):459–481, 2019. doi:10.3233/JCS-191287.
- [195] Lachlan Urquhart, Neelima Sailaja, and Derek McAuley. Realising the Right to Data Portability for the Domestic Internet of Things. *Personal and Ubiquitous Computing*, 22(2):317–332, 2018. doi:10.1007/s00779-017-1069-2.
- [196] US Census Bureau. 2017 American Community Survey, 2017. Accessed: 2020-03-16. URL: <https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t>.
- [197] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 26th ACM Conference on Computer and Communications Security, CCS '19*, pages 973–990, New York, NY, USA, 2019. ACM Press. doi:10.1145/3319535.3354212.
- [198] Pelayo Vallina, Álvaro Feal, Julien Gamba, Narseo Vallina-Rodríguez, and Antonio Fernández Anta.

- Tales from the Porn: A Comprehensive Privacy Analysis of the Web Porn Ecosystem. In *Proceedings of the 19th ACM SIGCOMM Internet Measurement Conference, IMC '19*, pages 245–258, New York, NY, USA, 2019. ACM Press. URL: <http://doi.acm.org/10.1145/3355369.3355583>, doi:10.1145/3355369.3355583.
- [199] Aysem Diker Vanberg and Mehmet Ünver. The Right to Data Portability in the GDPR and EU Competition Law: Odd Couple or Dynamic Duo? *European Journal of Law and Technology*, 8(1), 2017. URL: <http://ejlt.org/article/view/546>.
- [200] Evangelia Vanezi, Dimitrios Kouzapas, Georgia M. Kapitsaki, and Anna Philippou. Towards GDPR Compliant Software Design: A Formal Framework for Analyzing System Models. In *Proceedings of the 15th International Conference on Evaluation of Novel Approaches to Software Engineering, ENASE '20*, pages 135–162, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-40223-5_7.
- [201] Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy. FP-STALKER: Tracking Browser Fingerprint Evolutions. In *Pro-*

- ceedings of the 39th IEEE Symposium on Security and Privacy, S&P '18*, pages 728–741, Piscataway, NJ, USA, 2018. IEEE Computer Society. doi:10.1109/SP.2018.00008.
- [202] Giridhari Venkatadri, Alan Mislove, and Krishna P. Gummadi. Treads: Transparency-enhancing ads. In *Proceedings of the 17th ACM SIGCOMM Workshop on Hot Topics in Networks, HotNets '18*, pages 169–175, New York, NY, USA, 2018. ACM Press. doi:10.1145/3286062.3286089.
- [203] Natalija Vlajic, Marmara El Masri, Gianluigi M. Riva, Marguerite Barry, and Derek Doran. Online Tracking of Kids and Teens by Means of Invisible Images: COPPA vs. GDPR. In *Proceedings of the 2nd Workshop on Multimedia Privacy and Security, MPS '18*, pages 96–103, New York, NY, USA, 2018. ACM. doi:10.1145/3267357.3267370.
- [204] Xiaoyin Wang, Xue Qin, Mitra Bokaei Hosseini, Rocky Slavín, Travis D. Breau, and Jianwei Niu. GUILeak: Tracing Privacy Policy Claims on User Input Data for Android Applications. In *Proceedings of the 40th International Conference on Software Engineering, ICSE '18*, pages

- 37—47, New York, NY, USA, 2018. ACM Press.
doi:10.1145/3180155.3180196.
- [205] Lukas Weichselbaum, Michele Spagnuolo, Sebastian Lekies, and Artur Janc. CSP Is Dead, Long Live CSP! On the Insecurity of Whitelists and the Future of Content Security Policy. In *Proceedings of the 23rd ACM Conference on Computer and Communications Security, CCS '16*, pages 1376–1387, New York, NY, USA, 2016. ACM Press.
doi:10.1145/2976749.2978363.
- [206] Peng Weihong. HTTP Cookies—A Promising Technology. *Online Information Review*, 24(2):150–153, Jan 2000. doi:10.1108/14684520010330346.
- [207] Zachary Weinberg, Shinyoung Cho, Nicolas Christin, Vyas Sekar, and Phillipa Gill. How to Catch when Proxies Lie: Verifying the Physical Locations of Network Proxies with Active Geolocation. In *Proceedings of the 18th ACM SIGCOMM Internet Measurement Conference, IMC '18*, pages 203–217, New York, NY, USA, 2018. ACM Press.
doi:10.1145/3278532.3278551.

- [208] What Is My IP Address. See You Public IP Address, 2019. Accessed: 2020-03-16. URL: <https://whatismyipaddress.com/>.
- [209] Craig E. Wills and Can Tatar. Understanding What They Do with What They Know. In *Proceedings of the 11th ACM Workshop on Privacy in the Electronic Society, WPES '12*, pages 13–18, New York, NY, USA, 2012. ACM Press. doi:10.1145/2381966.2381969.
- [210] Janis Wong and Tristan Henderson. How Portable is Portable?: Exercising the GDPR’s Right to Data Portability. In *Proceedings of the 7th ACM International Joint Conference on Pervasive and Ubiquitous Computing and Wearable Computers, UbiComp '18*, pages 911–920, New York, NY, USA, 2018. ACM Press. doi:10.1145/3267305.3274152.
- [211] Janis Wong and Tristan Henderson. The Right to Data Portability in Practice: Exploring the Implications of the Technologically Neutral GDPR. *International Data Privacy Law*, 9(3):173–191, 07 2019. doi:10.1093/idpl/ipz008.
- [212] Qiang Xu, Rong Zheng, Walid Saad, and Zhu Han. Device Fingerprinting in Wireless Networks:

- Challenges and Opportunities. *IEEE Communications Surveys Tutorials*, 18(1):94–104, 2016. doi:10.1109/COMST.2015.2476338.
- [213] Shuai Yuan, Jun Wang, and Xiaoxue Zhao. Real-time Bidding for Online Advertising: Measurement and Analysis. In *Proceedings of the 7th Workshop on Data Mining for Online Advertising, ADKDD '13*, pages 1–8, New York, NY, USA, 2013. ACM Press. doi:10.1145/2501040.2501980.
- [214] Yong Yuan, Feiyue Wang, Juanjuan Li, and Rui Qin. A Survey on Real Time Bidding Advertising. In *Proceedings of the 9th IEEE International Conference on Service Operations and Logistics, and Informatics, SOLI '14*, pages 418–423, Piscataway, NJ, USA, 2014. IEEE Computer Society. doi:10.1109/SOLI.2014.6960761.
- [215] Jun Zhao, Ge Wang, Carys Dally, Petr Slovak, Julian Edbrooke-Childs, Max Van Kleek, and Nigel Shadbolt. “I Make up a Silly Name”: Understanding Children’s Perception of Privacy Risks Online. In *Proceedings of the 37th ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '19*, New York, NY, USA, 2019. ACM Press. doi:10.1145/3290605.3300336.

- [216] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. *Proceedings on Privacy Enhancing Technologies*, 2019(3):66–86, 2019. doi:10.2478/popets-2019-0037.
- [217] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven Bellovin, and Joel Reidenberg. Automated analysis of privacy requirements for mobile apps. In *Proceedings of the 23rd Symposium on Network and Distributed System Security*, NDSS '17, San Diego, California, USA, 2017. Internet Society. doi:10.14722/ndss.2016.23390.

LIST OF FIGURES

1.1	Overview of the contributions of this thesis.	13
1.2	Overview of the thesis structure.	14
3.1	Overview of prevalent website categories in our dataset.	60
3.2	Mean number of cookies set in our pre-study with the corresponding standard derivation	64

- 3.3 Overview of our measurement approach. First, we create the used website corpus, afterwards we create region-specific browser profiles and collect the websites to visit. In the final step, we visit the websites from the different regions and log the traffic. 66
- 3.4 Mean number of cookies used by each visited landing page and each respective subsite, by category of the visited website. To increase the readability, we capped the bars at 500. 1.8 % of sites had a higher number of cookies; this does not impact the computed values. 72
- 3.5 Classification of different cookies and the corresponding lifetime. Note the *logarithmic* scale. 74
- 3.6 Origins (left) and targets (right) of requests to services whose IP address is not mapped to an IP in an adequate country. 78
- 3.7 Example of an observed third party tree. The listed companies represent the companies operating the observed URLs. [C] illustrates the cookie setting parties. 93
- 3.8 Relative distribution of the measured third party tree depth split by the websites' categories. 97

5.3. *THESIS CONCLUSION* 309

- 3.9 Relative amount of cookies *not* set in each branch of the measured trees by category of the visited website (TLD+1). 99
- 3.10 Children included in only some of the branches (fluctuation) created by a specific TP within each visited site (grey) and across all sites (black). 103
- 3.11 Resulting branch depth of objects embedded by different companies (scaled for each individual company). 104
- 3.12 Amount of cookies set at each depth of the measured third party trees differentiated by cookies set in GDPR adequate countries (gray) and possibly conflicting countries (black). 106
- 3.13 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. 111
- 3.14 Overview of related work and how our work is distinct from it. 112

- 3.15 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. 116
- 3.16 Amount of domains and visited subsites used in our measurements (M#1–M#12). The red dots represent the amounts that occurred in M#1. 123
- 3.17 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 noticeable—the three most significant nodes are labeled. 124
- 3.18 Regression lines of our measurements including the pre-GDPR measurement (gray) and excluding it (black). The red dashes represents the confidence interval (99% confidence) for the prediction for the pre-GDPR measurement point based on all post-GDPR measurements. 129
- 3.19 Overview of the distribution of the measured degrees of all nodes (excluding isolated notes). 139

4.1	Shares of observed companies in terms of directly embedded (black) and actively syncing cookies (gray). The dotted lines show the share of the companies in our corpus.	161
4.2	Types and timings of the received responses.	177
4.3	Comparison of the workload to get access to personal data companies stored about a user.	182
4.4	Inferred <i>interest segments</i> provided by different companies (anonymized).	189
4.5	<i>Technical data</i> provided by different companies (anonymized).	190
4.6	<i>Tracking data</i> provided by different companies (anonymized).	191
4.7	Article recommendation including the company providing it (top right corner).	209
4.8	Traditional ad banner.	210
4.9	Evaluation of different aspects of the provided data.	219
4.10	Participants' view on the usefulness of profile categories.	219
4.11	Participants' ranking of profiles.	220

LIST OF TABLES

2.1 Overview of research conducted on subject access requests. The scale of each experiment is given with the respective success rates. An indicator is given if the SAR process (Proc.), the returned data (Data), the user authentication (Auth.), or malicious usage (Mal.) is analyzed. 39

- 2.2 Overview of literature that aims to measure the effects of the GDPR. The table shows if the study aimed to measure compliance of companies or the impact of the GDPR, the used technology to conduct the measurements, and the respective technologies that were measured. 47
- 3.1 General overview of our three measurement crawls. The number of visited websites and subsites with the corresponding number of observed TPs, cookie setting TPs (C TPs), and used cookies is shown. 70
- 3.2 Overview of previous work we tried to replicate (Rep.), the scale of the work (“LP” := landing page, “SB” := subsite), the results (Res.) of our replication, and if these experiments show different behavior in a vertical setup (Scales). 81
- 3.3 Overview of our measurements. For each measurement the number of visited domains, the visited number of subsites, and the observed third parties are given. 122

5.3. *THESIS CONCLUSION* 315

3.4 Overview of the measured graph structures (with and w/o isolated nodes) in terms of observed nodes (companies) and connections between them. The relative percentages refer to M#1. 125

3.5 Overview of connected components (CP) in the measured graphs (M#1–M#12) and the shift after the GDPR took effect. 132

3.6 Characteristics of our graphs w/o isolated nodes. 134

3.7 Betweenness centrality properties of our measured graphs and the change of the most central nodes over time. 138

4.1 The companies of our analysis corpus grouped by their respective business field(s). *AppNexus* (★) and *Adform* (♣) run two services and are therefore listed twice. *SpotX* is a subsidiary of the *RTL Group* (†). 160

4.2 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 171

4.3	Overview of the SAR process and responses for both rounds of inquiries.	181
4.4	Overview of the SAR process and responses for both rounds of inquiries.	186
4.5	Results of transparency tools analysis. ○: Does not apply. ●: Applies according to the privacy policy and data is provided. ◐: Applies according to the privacy policy, but no data is provided. ◑:= does not apply according to Privacy Policy, but data is provided. †: <i>Google</i> and <i>Facebook</i> only shows tracking data on their own platforms. <i>Twitter</i> 's way to provide sharing data did not work for us. <i>Sovrn</i> only shared pseudonyms of partners. ⚡: Our analyzed profile did not include this data but could include it. . . .	201
4.6	Privacy protection methods used by participants. Each participant used at least one the listed method or is listed among "None".	214
4.7	Participant Demographics. One participant preferred not to answer demographic questions (sex and age) and 4 preferred not to state their education level. ★: The census data does not account for non-binary individuals. ♣: The census data combines these categories.	215

APPENDIX A

SURVEYS

In this appendix, we present the questionnaires used in our human-centric approach (see Chapter 4).

A.1 User Survey Questionnaire

In our user study (see Section 4.2), we used the following questionnaire to evaluate user perception of current transparency tools provided by different ad-tech companies. All questions, excluding the consent form and open-ended questions (which can be left blank), offer a “*I prefer not*

to answer” option, which we omitted in the following for readability and space-saving. If not stated otherwise, we used: (5PL) 5 point Likert scales, (Y/N) yes, no, I do not know, or (OE) open-ended answer options (AO).

Question Group: Self Assessment

Q1: *Online advertisements or recommendations (e.g., for product items or articles) I see suit my interests.*

AO: 5PL from “Always” to “Never”.

Q2: *Rate your personal experience with ads or recommendations (e.g., for product items or articles) you see online.* AO: 5PL from “Very satisfied” to “Very dissatisfied”.

Q3: *Have you ever wondered why you see a specific ad or recommendation (e.g., for product items or articles) online?* AO: 5PL from “Always” to “Never”.

Q4: *Have you ever requested a copy of the personal data that a company has collected about you?* AO: Y/N

Q5: *I think having access to data ad companies collected about me is useful to better understand how they use my data.* AO: 5PL from “Strongly Agree” to “Strongly Disagree”.

Q6: *How do you rate your knowledge about online advertisement?* AO: 5PL from “Not knowledgeable” to “Very knowledgeable”.

Q7: *Which of the following statements regarding online advertisements do you think are true?* AO (multiple choice): (1) “Ad companies might share my data with their partners”, (2) “Ad companies have access to my full browsing history”, (3) “Ad companies collect personal data about me whenever they show me an ad”, (4) “Ad companies know which device I am using”, (5) “Ad companies have access to all products I bought online”, (6) “Ad companies learn my interests from my online actions”, and (7) “All statements are false.”

Q8: *Do you use any of the following mechanisms to protect your privacy online?* AO (multiple choice): (1) “I use a browser extension to block ads or to track (e.g., Ghostery, Adblock Plus, Privacy Badger, Disconnect)”, (2) “I browse in private mode (“incognito mode”) or use a VPN from time to time”, (3) “I delete cookies from my browser from time to time”, (4) “I opted out of online behavioral advertising with at least one company”, and (5) “None of the above”.

Q9: *Do you have any comments regarding your experience with online advertisements?* AO:

Open-ended

Question Group: Identifying Responsible Parties

Q10: *Take a look at the highlighted part of the following picture (red frame)[Figure 4.7]. If you wanted to understand on which data the specific ad or recommendation is based, whom would you ask?* AO

(multiple choice): (1) “The Review Experts”, (2) “ESPN.com”, (3) “Mansion Global”, (4) “ZaloTech”, (5) “Outbrain”, (6) “I do not know”.

Q11: *Take a look at the highlighted part of the following picture (red frame)[Figure 4.8]. If you wanted to understand on which data the specific ad or recommendation is based, whom would you ask?* AO

(multiple choice): (1) “Reddit.com”, (2) “The Outnet.com”, (3) “Google”, and (4) “I do not know”.

Question Group: Transparency Tool Assessment

On the next pages, you will see three different ways how companies make personal data accessible that they collected about someone. The categories are:

Technical data. Information automatically transmitted when you surf the web. *Tracking data.* Information on which websites the company tracked you. *Interest data.* User interests the company interfered from the collected data. We will provide you profiles of three different companies in each category (nine in total) that were collected about the same individual. Note: Companies may provide different information as they have different data sources

Q12-Q21: *Note: The following four questions were asked to each profile displayed in Figures 4.4, 4.5, and 4.6 (i.e., each question was asked 9 times). AO: (5PL) from “Strongly Agree” to “Strongly Disagree”.* (1) “This is the kind of information I expected to see.”, (2) “The website displays helpful information regarding personal data collected about me.”, (3) “I understand the presented information.”, and (4) “The information is presented clearly.”

Question Group: **General Transparency Questions**

Q22: *In general, do you think that companies provide all personal data they actually collected about you if you request them?* AO: Y/N.

Q23: *Prioritize which of the following information should be included if you request access to your personal data. By technical (raw) data we mean data displayed in the following images: [Figure 4.5] By tracking data we mean data displayed in the following images: [Figure 4.6] By interest data we mean data displayed in the following images: [Figure 4.4] You can skip this question if you prefer not to answer* AO: Ranking from 1 to 3 for each profile.

Q24: *In this survey, you saw personal data ad companies collected about a stranger, are you now interested in personal data collected about you?* AO: 5PL from “Very interested” to “Not at all interested”.

Q25: *Knowing what data ad companies collect about me, I would reconsider my online behavior.* AO: 5PL from “Strongly Agree” to “Strongly Disagree”.

Question Group: Improvement Suggestions

Q26: *From your point of view, what can ad companies do better to increase transparency in the online ad ecosystem?* AO: open-ended

Q27: *How can ad companies improve the presentation of collected personal data?* AO: open-ended

Question Group: Demographics

Q28: *How old are you?* AO: (1) “18-24”, (2) “25-34”, (3) “35-44”, (4) “45-54”, (5) “55-65”, and (6) “65 years or older”.

Q29: *How do you identify?* AO: (1) “Female”, (2) “Male”, and (4) “Non-binary”

Q30: *What is the highest degree or level of school you have completed? (If you’re currently enrolled in school, please indicate the highest degree you have received.)*

AO: (1) “Less than a high school diploma”, (2) “High school graduate”, (3) “Bachelor’s degree”, (4) “Master’s degree”, (5) “Professional degree”, and (6) “Doctorate degree”.

A.2 Company Survey Questionnaire

We used the following questionnaire to identify problems companies faced when implementing transparency tools (see Section 4.2.3). All questions, excluding the consent form and open-ended questions (which can be left blank), offer a “*I prefer not to answer*” answer option, which we omit in the following to increase readability and for space-saving.

Question Group: **General Questions**

Q1: *What is your role in your company?* AO:

(1)Legal / Privacy Management, (2) General Support / Helpdesk, (3) Data Protection Officer (DPO), (4) External / Consulting, and (5) Other(s) (please specify).

Q2: *Who is responsible for handling Subject Access Requests in your company?* AO: (1)Legal / Privacy Management, (2)General Support / Helpdesk, (3) Data Protection Officer (DPO), (4)External / Consulting, and (5) Other(s) (please specify).

Q3: (optional) *How high is the percentile amount of your data subjects that perform a Subject Access*

Request (e.g., “1 out of 10.000 data subjects” or “1%”)? AO: open-ended

Q4: Considering your expectations before the GDPR took effect: How often do you handle Subject Access Requests in your company? AO: 5PL from “Way more than expected.” to “Way less than expected. ”

Question Group: Development of the SAR process

Q5: Do you have a standardized process to handle Subject Access Requests in your company? AO: (1) Yes, there is a (semi)automated process, (2) No, each request is answered individually, or (3) I do not know.

Q6: Now, six months after the GDPR took effect, do you think it is necessary to change the way you handle Subject Access Requests? AO: (1) We already changed the way we handle such requests (since the GDPR took effect), (2) Yes, we plan to change the way we handle such requests, (3) No, but I think we should change the way we handle such requests, or (4) No.

Q7: Should there be a detailed guideline on how to handle Subject Access Requests? If so, who should provide it? AO: (multiple-choice) (1) Industry self-regulation (e.g., IBA or DAA), (2) Normative

regulation (e.g., in a standard/norm provided by ISO), (3) Legislative regulation (e.g., as an amendment of the GDPR), or (4) No more regulation is needed.

Q8: (optional) *What do you think were the most significant obstacles when designing your Subject Access Request process?* AO: open-ended

Question Group: Views on transparency

Q9: (optional) *Which benefits does the GDPR provide for your company?* AO: open-ended.

Q10:(optional) *To what extent are Subject Access Requests a useful tool for users to regain control of their data?* AO: open-ended

Q11: (optional) *Do you think that the GDPR provides any benefits to users when it comes to transparency in the online advertising industry?* AO: open-ended

Question Group: Demographics

Q12: *Is the headquarter of your company located in a country that is part of the European Union?* AO: (1) Yes, (2)No (please specify), or (3) I do not know.

Q13: *In which segments of the digital advertising ecosystem is your company active?* AO (multiple choice): (1) Agency / Agency Trading Desk, (2)

Targeting / Audience, (3) Data Management Platform / Data Provider, (4) Ad Exchange / Ad Server, (5) Verification & Privacy, (6) Demand Side Platform, (7) Ad network / Supply Side Platform, or (8) Other.

Q14: *How many people are employed at your company?* AO: (1) 1–50, (2) 51–100, (3) 101–250, (4) 251–500, (5) 501–1,000, or (6) > 1,000.

Q15: *How many employees work in your department?* AO: (1) 1, (2) 2–5, (3) 6–10, (4) 11–20, or (5) > 20.