

Tracking Mechanisms and the Effect of User Consent on the Web

Emmanouil Papadogiannakis

Thesis submitted in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science and Engineering

University of Crete
School of Sciences and Engineering
Computer Science Department
Voutes University Campus, 700 13 Heraklion, Crete, Greece

Thesis Advisor: Prof. *Evangelos Markatos*

This work has been performed at the University of Crete, School of Sciences and Engineering, Computer Science Department.

The work has been supported by the Foundation for Research and Technology - Hellas (FORTH), Institute of Computer Science (ICS).


UNIVERSITY OF CRETE
COMPUTER SCIENCE DEPARTMENT

Tracking Mechanisms and the Effect of User Consent on the Web


Thesis submitted by
Emmanouil Papadogiannakis
in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science

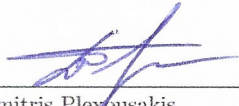
THESIS APPROVAL

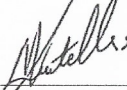
Author:


Emmanouil Papadogiannakis


Committee approvals:


Evangelos Markatos
Professor, Thesis Supervisor


Dimitris Plexousakis
Professor, Committee Member


Nicolas Kourtellis
Senior Research Scientist, Committee Member

20/01/2022


Panagiotis Papadopoulos
Security Researcher, Committee Member

Departmental approval:


Polyvios Pratikakis
Associate Professor, Director of Graduate Studies

Heraklion, January 2022

Tracking Mechanisms and the Effect of User Consent on the Web

Abstract

During the past few years, mostly as a result of legislation such as the GDPR and the CCPA, websites have started presenting users with consent banners. These banners are web forms where users can state their preference regarding data processing purposes and declare which cookies they would like to accept. Although requesting consent before storing any identifiable information is a good start towards respecting the user privacy, previous research has shown that websites do not always respect user choices. Furthermore, considering not only the ever decreasing reliance of trackers on cookies, but also the actions browser vendors take by blocking or restricting third-party cookies, we anticipate a world where stateless tracking emerges, either because trackers or websites do not use cookies, or because users simply refuse to accept any.

In this thesis, we explore whether websites use more persistent and sophisticated forms of tracking in order to track users who explicitly denied consent. Such forms of tracking include first-party ID leaking, third-party ID synchronization and browser fingerprinting. Using a novel web crawler, we examined the consent banners of over 27,000 websites and our results suggest that websites do use such modern forms of tracking even before users had the opportunity to register their consent choice. To add insult to injury, when users choose to raise their voice, deny consent and reject all cookies, user tracking only intensifies. We measured that aggressive tracking takes place before users had the opportunity to make a selection in the consent banner, with more than 75% of tracking activities happening when users chose to deny consent. Consequently, we conclude that user choices play very little role with respect to sophisticated tracking mechanisms.

Μηχανισμοί Ιχνηλάτησης και η Επίδραση της Συναίνεσης του Χρήστη στο Διαδίκτυο

Περίληψη

Τα τελευταία χρόνια, κυρίως λόγω νομοθεσίας όπως είναι το GDPR και το CCPA, οι ιστοσελίδες έχουν ξεκινήσει να παρουσιάζουν παράθυρα συναίνεσης στους χρήστες. Αυτά τα παράθυρα είναι διαδικτυακές φόρμες στις οποίες οι χρήστες μπορούν να δηλώσουν τις προτιμήσεις τους σχετικά με την επεξεργασία των δεδομένων τους αλλά και ποια cookies επιθυμούν να αποδεχτούν. Παρόλο που το να ζητάνε συναίνεση προτού αποθηκεύσουν οποιαδήποτε αναγνωρίσιμη πληροφορία είναι μια καλή αρχή για τον σεβασμό της ιδιωτικότητας των χρηστών, προηγούμενη έρευνα έχει δείξει ότι οι ιστοσελίδες δεν σέβονται πάντα τις επιλογές των χρηστών τους. Επιπλέον, λαμβάνοντας υπόψη όχι μόνο την φθίνουσα εξάρτηση των ιχνηλατών (trackers) στα cookies, αλλά και τις ενέργειες των προμηθευτών προγραμμάτων περιήγησης για την παρεμπόδιση cookies τρίτων, προβλέπουμε έναν κόσμο όπου θα κυριαρχήσει η ιχνηλάτηση χωρίς μνήμη, είτε επειδή οι ιχνηλάτες ή οι ιστοσελίδες δεν χρησιμοποιούν cookies, είτε απλά επειδή οι χρήστες αρνούνται να τα αποδεχτούν.

Σε αυτή την εργασία, εξερευνούμε εάν οι ιστοσελίδες χρησιμοποιούν πιο επίμονες και προχωρημένες μορφές ιχνηλάτησης με σκοπό να ανιχνεύουν χρήστες οι οποίοι αρνούνται ρητά συναίνεση. Τέτοιες μορφές ιχνηλάτησης συμπεριλαμβάνουν first-party ID leaking, third-party ID synchronization και browser fingerprinting. Χρησιμοποιώντας έναν καινοτόμο crawler διαδικτύου, εξετάσαμε τις φόρμες συναίνεσης περισσότερων από 27,000 ιστοσελίδων και τα αποτελέσματά μας δείχνουν ότι οι ιστοσελίδες όντως χρησιμοποιούν τέτοιες μοντέρνες μορφές ιχνηλάτησης πριν ακόμα οι χρήστες έχουν την ευκαιρία να καταγράψουν την επιλογή τους. Ακόμη χειρότερα, όταν οι χρήστες επιλέγουν να υψώσουν τη φωνή τους, να αρνηθούν συναίνεση και να απορρίψουν όλα τα cookies, ο εντοπισμός εντείνεται. Μετρήσαμε την επιθετική ιχνηλάτηση να λαμβάνει χώρα προτού οι χρήστες έχουν την ευκαιρία να πάρουν μια απόφαση στις φόρμες συναίνεσης με περισσότερο από 75% των ενεργειών ιχνηλάτησης να γίνονται όταν οι χρήστες επιλέγουν να αρνηθούν συναίνεση. Ως αποτέλεσμα, καταλήγουμε ότι οι επιλογές των χρηστών παίζουν πολύ μικρό ρόλο όσον αφορά τους προχωρημένους μηχανισμούς ιχνηλάτησης.

Acknowledgements

I would like to thank all the people that helped me and influenced me during my studies. First and foremost, I am more than grateful to my supervisor Prof. Evangelos Markatos for his guidance, continuous support and for the chance to be part of his research team. Your immense knowledge and insightful feedback has pushed me to become a better researcher and your passion for research has shaped my mentality. You have been an incredible teacher to me.

I need to also express my sincere gratitude to Dr. Nicolas Kourtellis and to Dr. Panagiotis Papadopoulos for their help, guidance and continuous support. We spent numerous hours working together, devising methodologies and discussing results. Your advice and mentorship made this journey beautiful. Most of all, I would like to thank you for trying to teach me what it means to be a researcher.

I wish to show my appreciation to all past and present members of the Distributed Computing Systems and Cybersecurity laboratory at ICS-FORTH. Their technical and moral support made my working days both productive and enjoyable. Also, I would like to thank all the professors, faculty and teaching assistants of the Computer Science Department of the University of Crete.

Last but not least, I would like to thank, from the bottom of my heart, my parents Angelos and Maria and my sister Eva for their unconditional love, support and encouragement through my life. You are always there for me and I wouldn't have made it this far without you...

This work was performed at the **Distributed Computing Systems and Cybersecurity** laboratory, **Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas (FORTH)** and has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 830929 (CyberSec4Europe) and No 786669 (REACT). The results presented in this thesis reflect only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

to my family

Contents

Table of Contents	i
List of Figures	v
List of Tables	iii
1 Introduction	1
1.1 Contributions	3
1.2 Outline	4
2 Background	5
2.1 Cookies	5
2.2 ID Sharing	5
2.3 Browser Fingerprinting	7
2.4 Legislation and Consent	8
2.5 Consent Management Platforms	10
3 Methodology	11
3.1 Crawling Methodology	11
3.2 Data Description	12
3.3 Detecting ID Sharing	13
3.4 Detecting Browser Fingerprinting	15
4 Analysis of Consent	17
4.1 Consent and Third-Parties	17
4.2 Sharing User IDs with Third-Parties	18
4.3 Third-party ID synchronization	21
4.4 Browser Fingerprinting	22
4.5 Does website popularity matter?	23
4.6 Does the hosting country matter?	26
5 Ineffective Consent: Edge cases	29
6 Related Work	31

7	Discussion	35
7.1	GDPR compliance	35
7.2	Outbound Information	36
7.3	Edge Cases	36
7.4	Methodology	36
7.5	Limitations	37
8	Conclusion	39
9	Ethical Considerations	41
	Bibliography	43

List of Figures

2.1	Example of an ID synchronization operation. Two entities link different IDs they have assigned to the same user.	6
2.2	The Canvas Fingerprinting process as part of the browser fingerprinting methodology used by popular libraries. The website can extract an almost unique fingerprint of the user's browser.	7
2.3	Example of a consent banner where users can state their preference regarding the processing of their personal data. Users can also choose to get a detailed view of the third-parties, which collect data or with which data is shared.	9
3.1	High level overview of our crawling methodology. We use Puppeteer to instrument a web browser and automatically visit websites. The Chrome Profiler is a built-in tool used to record and analyze runtime performance by collecting callsite information and execution statistics. The Cookie Database stores all cookies set by various domains. The Consent-O-Matic tool is loaded on browser startup as an extension to handle consent banners. Whenever a request is issued or a response is received, the event dispatcher emits the appropriate event, which is handled by our puppeteer-based crawler.	12
3.2	Example of a first-party identifier being leaked to a third-party. A user ID assigned by <i>trivago.co.id</i> is being shared with Google.	14
3.3	Snippet of the popular fingerprinting library, FingerprintJS.	16
4.1	Number of third-parties running on the website during the three different types of visits (i.e., different consent actions). Surprisingly, for the median website, in the Reject All case there were more (i.e., 17) third-parties running than in the No Action case (i.e., 16).	18
4.2	Average number of unique third-parties learning a user identifier. A user's browser leaks first-party identifiers to 2.14 third-parties, on average, even before the user accepted or rejected consent.	19
4.3	Mean number of third-parties involved in first-party ID leaking per website rank.	24
4.4	Unique third-parties in ID leaking per website rank (normalized).	24
4.5	Unique third-parties in ID leaking per website rank (linear regression).	24

4.6	Unique third-parties involved in ID synchronization per website rank.	25
4.7	Unique third-parties involved in ID synchronization per website rank (normalized).	25
4.8	Unique third-parties involved in ID synchronization per website rank (linear regression).	25
4.9	third-party ID synchronization as a function of the website's popularity. (a) Average number of unique third-parties involved in synchronizations per rank range of the website. (b) This figure plots the same information as Fig. 4.6, with the exception that all Accept All values are normalized to 100%. (c) In this figure exception that Reject All and No Action points have been fitted with a straight line. The line suggests an increasing trend implying that less popular sites are more aggressive at disregarding user choices.	25
4.10	Number of unique third-parties learning a first-party identifier as a function of the top-level domain per country code.	26
4.11	Normalized number of unique third-parties learning a first-party identifier. This figure plots the same information as Figure 4.10, with the difference that the max value (Accept All) is normalized to 100%. This enables us to compare websites that have different magnitudes of leakage.	27
4.12	Normalized number of unique third-parties engaged in third-party ID synchronization.	27

List of Tables

3.1	Summary of our crawled dataset.	13
4.1	Number of websites detected (i) leaking their first-party user identifiers and (ii) having third-parties that perform synchronizations of user IDs.	19
4.2	Top 5 third-parties that learn the highest number of first-party identifiers per consent action in our dataset. For each party we compute the percentage of leaked first-party identifiers that they learn. . . .	20
4.3	Top 5 third-parties with highest number of third-party synchronizations per consent action in our dataset.	21
4.4	Websites performing browser fingerprinting.	23

Chapter 1

Introduction

Over the past few years, an increasing concern to protect user privacy has been observed, with the protection of data of users across the world being the main focus. Specifically in Europe, this is evident in the General Data Protection Regulation (GDPR) which was adopted on April 2016 and came into force on May 2018. This regulation focused on citizens of the European Union trying to enable them to have better control of their personal data. However, it is important to notice that the GDPR has a global effect since it affects all websites that process data of European citizens, even if these websites are based outside the European Union. The main difference of this regulation compared to previous legislation is that it includes significant fines for companies which collect user data without the user's consent or some other legal basis.

As a result, several companies, and their associated websites, have started asking their visitors for their consent, before collecting and processing their data. Beyond that, some websites have even taken extreme measures like blocking users and refusing services to visitors from the European countries [68, 28]. A careful reader of the GDPR, can surmise that the regulation is not about cookies alone, but instead about proper handling of any type of personal data. Any information that might be linked to an identifiable individual is considered personal data. This definition covers not only online identifiers (e.g., device's MAC address, IP address, or an advertising user identifier stored in a cookie), but also less specific features like the combination of browser characteristics.

Websites usually collect consent through the use of consent banners, which ask users for their preference regarding data processing and may also provide some options. Indeed, users may be given the choice to accept all cookies, to accept only some, or even to reject them all. The choice is entirely up to the user, and the correct implementation of this choice is the responsibility of the website. Due to the technical and legal complexity of this task, during the last few years, we have witnessed the advent of Consent Management Providers, third-party services which are tasked with providing users with the appropriate options regarding data processing, and then, properly collecting and storing their consent.

Although this sounds completely legal and fully straightforward, deviations and discrepancies between what the users select and what is in fact registered in the website have been reported in literature [27, 18, 67, 45, 75]. For example, users may provide a negative response (i.e., reject all cookies), but cookie banners may register a positive one (i.e., accept all cookies), or the cookie banners may register a positive response even before the users has had the opportunity to provide any choice [45]. Additionally, websites have the legal right to declare some cookies as absolutely necessary for their operation (e.g., for the page to be delivered) or due to legitimate interest (e.g., to improve the product or the provided service). Such cookies can not be rejected by the users, thus, users often do not really have the option to reject *all* cookies.

In this thesis, we explore a slightly different question:

*If a user does not provide consent, or chooses to **reject** all cookies, do websites use other forms of tracking to track this user? If so, what are these forms of tracking, and what is the extent of this tracking?*

In the Web ecosystem, cookies are often used to store identifying information for a given user. Indeed, the digital advertising industry depends on personalization and targeted ads. Previous work has focused on cookies and compliance of cookie processing with the GDPR and similar legislation (e.g., [67, 13]). However, recent policies and regulations from browser vendors and government bodies try to control the exposure of this identifying information to third-parties [11]. In fact, almost all modern browser vendors have taken action to impede the use of third-party cookies [78, 72, 7]. These policies restrain the ad and tracking industry that relies on re-identifying a user for long periods to serve more targeted ads.

Considering the (i) ever less reliance of third-party trackers on non-permanent, erasable state-like cookies [37], (ii) recent advances of browser vendors against third-party cookies [10, 77] and (iii) the introduction of new technologies in the advertising ecosystem [16], it is apparent that the need for identifying how websites treat user consent in case of stateless (i.e., cookie-less) tracking is more than timely and urgent. We address this need and try to fill this exact gap in our understanding, by being the first to investigate what is the GDPR compliance across the Web in the case where websites and trackers do not use cookies, or users do not accept cookies.

Sadly, our results suggest that even when users deny consent, websites do use other forms of tracking to track users, and process their personal data, in violation of legislation, such as the GDPR. Such forms of tracking include *first-party ID leaking*, *third-party ID synchronization* and *browser fingerprinting*. One might think that the synchronization or leak of identifiers is a form of tracking using cookies, however, that is not entirely true. Although ID synchronization might use values stored in cookies, passing such values around is done in an unorthodox manner completely different from the way cookies are used. Particularly, first-party ID leaking and third-party ID synchronization are used to pass an identifier as an “argument” in network traffic towards a party different from the one that

assigned it in the first place. In fact, according to past studies [58, 59, 24], Web entities may share the identifiers they have assigned to users and help third-parties re-identify users or even create universal IDs.

On the other hand, browser fingerprinting [20, 1] uses elaborate techniques to uniquely identify users through characteristics of their devices. This technique focuses on characteristics which can be easily found and are readily accessible by a website. Such characteristics may include screen resolution, rendering attributes, browser fonts, installed plugins, etc. [17, 46, 51, 57]. Combining several of these characteristics can lead to signatures that uniquely identify users.

Although these cases of user identification are considered “personal data processing” according to GDPR and ePrivacy [9] regulations, and must be visible to users, they often do not appear in request forms of consent managers deployed by modern websites. In this study, we highlight the lack of transparency and the disparity regarding user consent and the sophisticated tracking mechanisms that websites deploy.

1.1 Contributions

In this thesis, we make the following main contributions:

1. We propose a novel and fully automated method for detecting browser fingerprinting in websites by utilizing the Chromium Profiler.
2. We perform a large-scale analysis by crawling close to one million websites and record how they track users using sophisticated tracking mechanisms, such as first-party ID leaking, third-party ID synchronization and browser fingerprinting, as a function of user choices in consent banners.
3. We find that (1) Websites embedded with ID synchronizing third-parties force browsers to engage in several ID synchronizations even before users had a chance to accept or deny consent; (2) More than 75% of leaks happen despite the fact that users have chosen to reject all cookies; (3) Less popular websites are more likely to disregard users’ consent choices and engage in first-party ID leaking and third-party ID synchronization; (4) Browsers leak more information when users choose to reject all cookies than when they choose to take no action at all; (5) Our analysis of tracking per country code reveals significant discrepancies across EU countries.
4. We make our crawling and analysis tool publicly available [56] to support further research on this topic.
5. Our methodology can be transformed into an auditing tool for regulators, stakeholders and privacy-policy makers, for verifying compliance with GDPR and users’ privacy rights.

1.2 Outline

The rest of this thesis is organized as follows. First, we describe some basic background knowledge concerning advanced and sophisticated tracking mechanisms, as well as, the legislation related to our study. We then present the design of the system we developed to study these tracking mechanisms based on the user's consent action along with the algorithms we used to detect consent violations. Then, we perform an offline analysis on the behavior of websites and the consent given by the user. We study the interaction with third-party trackers, the extent of browser fingerprinting and identifier leakage, as well as, the entities which are involved in first-party ID leaking and third-party ID synchronization, followed by a study of how the popularity and the hosting country of websites affect its behavior with regards to tracking mechanisms. Next, we present some extreme cases of ineffective consent. We conclude with providing previous related work, as well as, a discussion on the findings of this thesis.

Chapter 2

Background

In this chapter, we present some technical background related to tracking in the advertising ecosystem, as well as, a short overview of relevant legislation.

2.1 Cookies

Cookies are a simple mechanism that allows for stateful browsing over the stateless HTTP(S) protocol. Cookies are defined by a name, a value, an expiration date, as well as the domain they are associated with. Using cookies, servers and web applications can instruct browsers to store data in the client side. Whenever a request is issued towards a domain, the browser will automatically attach all cookies related to this domain.

Historically, cookies have been used to keep track of user sessions, store user information, such as preferences, and to facilitate complex web applications. However, over the years, cookies have been used to track users and their behavior across the web. Often, unique identifiers are assigned to the visitor and stored in cookies in order for first-parties to re-identify users across visits and third-parties to track users across multiple websites. Consequently, various Web entities are able to build profiles based on the users' browsing behavior and learn about their online activity. These profiles can be later centralized in Data Management Platforms [19], sold by data brokers [3], or used by advertisers to bid in ad auctions [55] or for ad-retargeting [35] and cross-device tracking [70].

2.2 ID Sharing

As mentioned in Chapter 2.1, whenever users visit websites, various parties assign identifiers in order to be able to re-identify users across multiple visits or multiple websites. However, since cookies might contain sensitive information, they can only be sent to the domain that set them or any parent domain. In order for the Web entities (e.g., publishers, analytics, data brokers, advertisers) to be able to track users efficiently and consolidate user profiles, they need to link (i.e., synchronize)

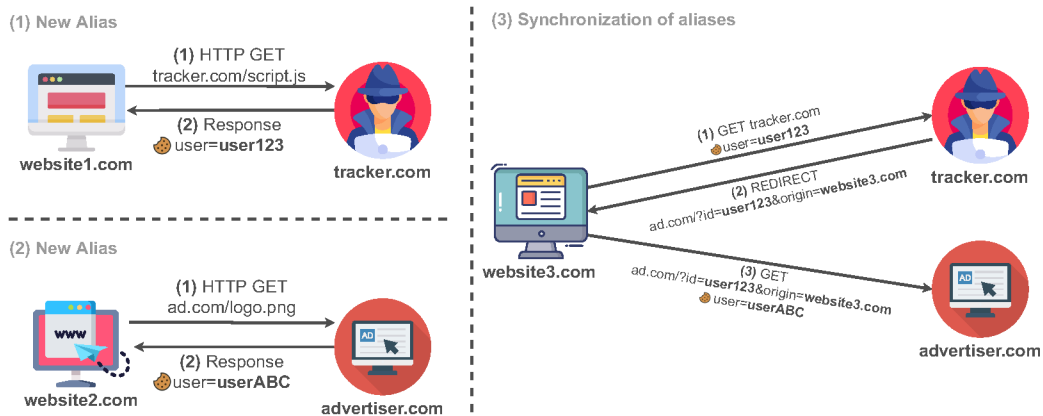


Figure 2.1: Example of an ID synchronization operation. Two entities link different IDs they have assigned to the same user.

together the different aliases that each entity has assigned to the same user. This would reveal that the user entity A knows as `userABC` is the same user that entity B knows as `user123`.

Figure 2.1 illustrates an example of how ID synchronization takes place. Assume a user browsing `website1.com` and `website2.com`, in which there are third-parties like `tracker.com` and `advertiser.com` respectively. Consequently, these two third-parties have the chance to assign an alias to the user and re-identify them in the future. From this point forward, `tracker.com` knows the user with the identifier `user123`, and `advertiser.com` knows the same user with the ID `userABC`. Next, assume that the user lands on `website3.com`, which includes some JavaScript code from `tracker.com`. This code will instruct the browser to issue a GET request to `tracker.com` (step 1), which responds back with a REDIRECT request (step 2), instructing the user’s browser to issue another request to its collaborator `advertiser.com` this time, using a specifically crafted URL (step 3) where the alias it uses (i.e., `user123`) is also attached to the issued request. When `advertiser.com` receives this request from the user it knows as `userABC`, it learns that this, in fact, is the same user that `tracker.com` knows as `user123`. In this figure, we illustrate a simple information flow from `tracker.com` to `advertiser.com`. However, the advertiser can readily share its own alias by using exactly the same redirection mechanism. As a result, both entities will be able to learn both aliases, which have been assigned to the same user. Altogether, this mechanism allows multiple third-party entities to join the different aliases (e.g., user IDs, device IDs etc.) a single user has on the Web.

In this thesis, we study two types of identifier sharing: (i) *first-party ID leaking*, where a first-party alias (e.g., a cookie or device ID) is leaked from the visited website to different third-parties via network traffic, and (ii) *third-party ID synchronization*, where third-parties link together the different third-party aliases they use for the same users.

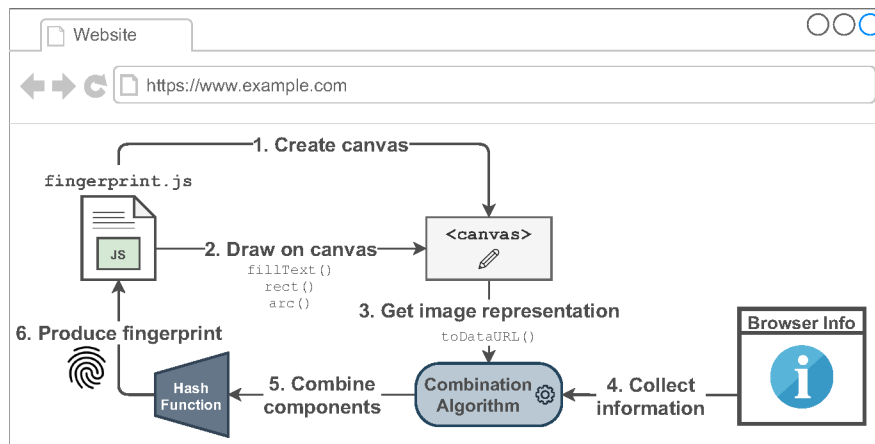


Figure 2.2: The Canvas Fingerprinting process as part of the browser fingerprinting methodology used by popular libraries. The website can extract an almost unique fingerprint of the user’s browser.

2.3 Browser Fingerprinting

Browser Fingerprinting is a sophisticated set of techniques, which can be used to uniquely identify browser instances without storing any information on the user side (*stateless*). It can be used to detect malicious users that create multiple accounts in social networking services, or even stop deceitful orders in e-commerce platforms. However, this technique can be abused by privacy-violating websites and, therefore, track users across websites, or even de-anonymize private sessions. In fact, previous work [48, 41] has demonstrated that this technique provides sufficient bits of entropy to effectively track users, even through the usage of the Tor Browser.

One of the most prevalent and stealthy such fingerprinting techniques is Canvas Fingerprinting, named after the HTML canvas element, which was introduced in the latest version (i.e., HTML5). A canvas element provides the required functionality for drawing graphics using client-side code. Moreover, canvas fingerprinting relies on WebGL, a cross-platform JavaScript API that enables developers to render advanced graphics using shaders. As a result, developers have access to rendering functionality, which is performed in a GPU, however, in an HTML context via the canvas element.

Tracking techniques need to be transparent to users to avoid raising suspicion or harm the user experience. As such, browser fingerprinting can be performed in minimum time on any browser that supports JavaScript by using invisible HTML elements and without requiring any privileges or permission from the user. Consequently, even privacy-aware advanced users that block cookies can be tracked. Furthermore, browser fingerprinting is difficult to prevent because it relies on native functionality, provided by modern browsers. As a result, users need to either

disable JavaScript, or resort to external browser extensions. These extensions usually add random noise to some built-in functions, making the fingerprint different each time a website attempts to (re)identify a user [40, 50]. In addition to this, modern browser vendors employ new techniques to protect their users from being tracked. Safari provides simplified system information to scripts [30], Firefox provides the Enhanced Tracking Protection option which block fingerprinters [49] and Brave randomizes “fingerprintable” values in order to confuse fingerprinting scripts [6].

Figure 2.2 demonstrates the process of canvas fingerprinting as part of the browser fingerprinting mechanism. Assume (i) a website that contains the fingerprinting code and (ii) a browser instance that can execute JavaScript code. As a first step, the fingerprinting script creates a canvas element using the built-in interface provided by almost all modern browsers. Next, the script renders some 2D graphics and text on the canvas. Usually, the text that is drawn is a “pangram”, an alphanumeric value that contains all the letters of the English alphabet in order to increase the number of entropy bits. Different font sizes and font families result in a slightly different text that can affect the final fingerprint. As a next step, the fingerprinting script extracts the content of the canvas and inspects its pixel values (step 3). This is achieved using the native method `toDataURL()`, provided by the canvas object, that returns the Base64 encoding of the content rendered inside the canvas. Based on various factors, including the fonts which are installed on the user’s machine, the version of OpenGL and the browser’s rendering engine, this string can be sufficiently different per browser instance.

Then, the script combines this canvas fingerprint with other information, which can be used as an additional source of entropy (step 4). This information includes, among others, the host operating system and timezone, its screen resolution, installed plugins, preferred language set in the browser and number of logical processors available on the host. Additionally, using more sophisticated algorithms, the script can also extract information using the browser’s audio context [64], WebRTC support [25], or even the Battery API [53]. The output of the combination algorithm is a long alphanumeric value that almost uniquely identifies the specific browser instance (step 5). Finally, this value is hashed, to produce a fingerprint for this specific browser instance (step 6) and is usually sent across the network, or even stored as a cookie. Popular fingerprinting libraries claim that they can achieve a fingerprinting accuracy of over 99% in modern browsers [33].

2.4 Legislation and Consent

The *General Data Protection Regulation* (GDPR) [62] is an initiative by the European Union that specifies the circumstances under which personal data may be processed. Coming into force on 25 May 2018, its goal was to establish and extend the rights that individuals have on their personal data. Even though the GDPR is an EU legislation, its impact is global since it applies even to entities, located

This website uses cookies

We use cookies to personalize content and ads, to provide social media features and to analyze our traffic. We also share information about your use of our site with our social media, advertising and analytics partners who may combine it with other information that you've provided to them or that they've collected from your use of their services.

Please choose whether this site may use cookies or related technologies such as web beacons, pixel tags, etc. You can learn more about how this site uses Cookies and how your personal data may be processed by [managing your settings](#) and reading our [privacy policy](#).

Necessary	Preferences	Statistics	Marketing
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Details ▼
Reject All
Accept All

Figure 2.3: Example of a consent banner where users can state their preference regarding the processing of their personal data. Users can also choose to get a detailed view of the third-parties, which collect data or with which data is shared.

outside the EU, that process personal information of EU residents. Failing to comply to the regulation results in large fines of up to 20 million Euros or 4% of a company's annual global turnover (whichever is the largest).

Personal data is defined in the GDPR as any information related to an identifiable natural person. Additionally, the GDPR does not differentiate between first-party and third-party identification and it also concerns pseudonymous data that can be attributed to a specific person. Article 6, contains six specific legal bases for the lawful processing of personal data. According to that, Web entities can establish this legal basis by obtaining the user's consent to the processing of personal data for one or more specific purposes. Moreover, articles 4 and 7 define that a valid consent is freely given, specific, informed, unambiguous, readable, accessible and revocable.

The ePrivacy Directive (2002/58/EC) [61] is an EU directive that requires consent before data is stored in user devices. An exemption is set, nonetheless, for cases that storing information is "strictly necessary" to perform the offered service. However, before the GDPR, this directive was usually implemented by simply displaying a message about the use of cookies, without providing an option to the user.

Similar to the GDPR, the *California Consumer Privacy Act* (CCPA) [44] gives consumers more control over their personal information by granting them the right to know about the personal information a business collects. The CCPA came into effect on January 2020 in the United States of America and explicitly states that a notice list the categories of personal information that the business collects and that this notice must be provided at or before the point at which the business collects the information. Additionally, users have the right to opt-out of the sale of their personal information and to request that their personal data is deleted.

2.5 Consent Management Platforms

When privacy regulations came into effect around the world, both publishers (i.e., first-parties) and advertisers or trackers (i.e., third-parties) were forced to properly collect and communicate user consent before processing user data. This process required technical knowledge in order to implement the correct systems that comply with the respective privacy law that applies to the respective user. To that extent, the Interactive Advertising Bureau Europe (IAB) published the Transparency and Consent Framework (TCF) [23] in order to help web entities comply with legislation. In this document, *Consent Management Platforms* (CMPs) were introduced. CMPs are third-party services that provide the needed functionality to publishers and are responsible for collecting and storing the user's consent. Usually, this is performed by providing the appropriate user-interface and corresponding functionality of consent banners, i.e., web forms where users can state their preferences regarding data processing. Figure 2.3 illustrates an example of a consent banner. Additionally, CMPs are responsible for implementing the necessary functionality and mechanisms in order for third-parties to be able to learn this consent. All CMPs that are compliant with the framework are registered in a public list [22].

Chapter 3

Methodology

In this chapter, we present the methodology we followed in order to crawl websites and collect data. Then, we describe the dataset we formed and the techniques and algorithms we followed to detect the tracking mechanisms described in Chapters 2.2 and 2.3.

3.1 Crawling Methodology

To investigate the effect of the different options a user is provided with while visiting websites with a consent form, we leverage the Consent-O-matic tool [26]. Consent-O-Matic is a state-of-the-art browser extension, which is able to automatically detect and handle consent banners of popular CMPs. Particularly, whenever the extension detects a Consent Management Platform (CMP), it logs the respective information (e.g., vendor, encoding, IDs) and automatically interacts with the consent banner. Additionally, the extension can be configured to either accept or reject the different categories of data processing purposes.

In addition to this, we develop a Web crawler, based on the *Puppeteer* framework [34], that instruments instances of the Chromium browser. By utilizing the Consent-O-Matic extension, the browser can automatically perform one of the following three actions when a consent form is detected:

1. **Accept All:** grant consent for all data processing purposes to all third-parties residing in the visited website.
2. **Reject All:** deny consent for all data processing purposes to all third-parties residing in the visited website.
3. **No Action:** avoid interacting with the form in any way.

By using our instrumented browser, we crawl with clean state the landing page of the top 850,000 most popular websites of the Tranco list [63]. This list aggregates the ranks from the lists provided by Alexa, Umbrella, and Majestic

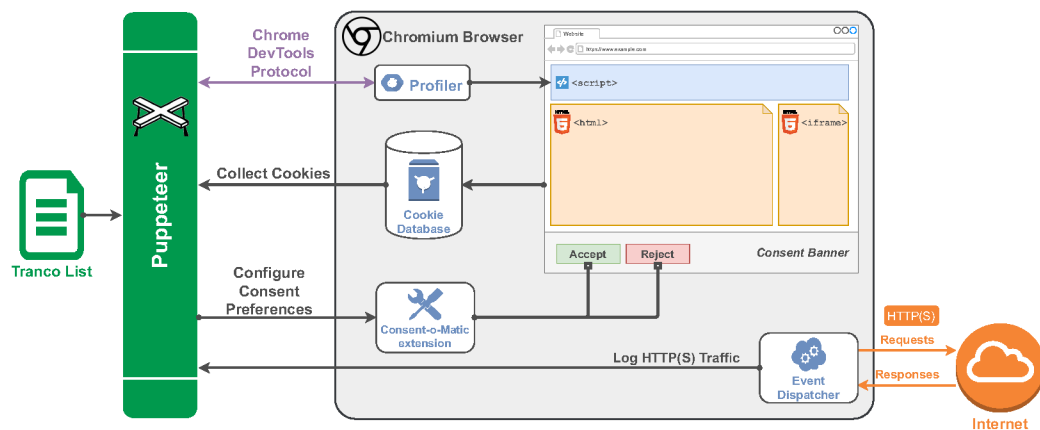


Figure 3.1: High level overview of our crawling methodology. We use Puppeteer to instrument a web browser and automatically visit websites. The Chrome Profiler is a built-in tool used to record and analyze run-time performance by collecting call-site information and execution statistics. The Cookie Database stores all cookies set by various domains. The Consent-O-Matic tool is loaded on browser startup as an extension to handle consent banners. Whenever a request is issued or a response is received, the event dispatcher emits the appropriate event, which is handled by our puppeteer-based crawler.

from 29 July 2020 to 27 August 2020¹. Whenever the Consent-O-Matic extensions detects a consent banner, we crawl the given website three times, one for each of the different consent actions. For each visit we store the HTML of the website, a cookie-jar for both first-party and third-party cookies, HTTP(S) requests and responses, JavaScript function calls and information about the detected CMP. Most importantly, we capture all application-level network traffic passively, via the emitted Chrome events without mutating or intercepting them. This ensures that the behavior of the website is not affected by our crawler. An overview of our crawling methodology is illustrated in Figure 3.1. The implementation of our crawler is publicly available as a standalone tool [56]. Ethical aspects regarding the crawling process and collected data are addressed in Chapter 9.

3.2 Data Description

Overall, our crawler was located in the EU and visited 850K websites from August 28th 2020 to September 17th 2020. The Consent-O-Matic extension detected 27,953 websites with a consent banner (or 4.44% of the successfully visited websites) and our crawler collected a total of 108 GB of data for these websites. This finding is inline with related work which reports detection rates of 3% [67] and 6.2% [45].

¹<https://tranco-list.eu/list/Q274/full>

Table 3.1: Summary of our crawled dataset.

Description	Volume	% of total
Initial set of websites	850,000	
Websites that errored	219,098	25.78%
Websites that were filtered out	2,689	0.32%
Total websites correctly parsed	628,213	73.90%
Websites with a CMP	27,953	3.29%
Websites with a CMP and no error in all three consent actions	27,180	3.20%

Low detection rates are attributed to the fact that the Consent-O-Matic extension is only able to detect popular consent banners following a specific standard. Designing a detection tool with better accuracy is a very challenging task due to the heterogeneity of the various existing consent management libraries and custom solutions adopted by websites. This problem, though interesting, is considered as out of scope for this work and is left for future research. However, we performed manual inspection of websites and verified that the Consent-O-Matic extension was able to properly handle the appropriate banners. Specifically, we found that for popular Consent Management Platforms, such as *quantcast* and *cookieLab*, the extension was able to properly handle 100% of the evaluated websites.

Moreover, crawls failed at 25.78% of the initial set of websites due to timeouts or site inaccessibility (i.e., site did not serve EU-based users). Additionally, we filtered out websites that explicitly defined in their `robots.txt` file that they do not allow crawling, as well as, some popular pornographic websites due to internal organization policies. Table 3.1 summarizes the collected dataset.

3.3 Detecting ID Sharing

We perform an offline analysis on the collected data to detect ID sharing operations. Specifically, we examine all application-level network traffic and search for requests that contain values that uniquely identify users. Specifically, for HTTP(S) GET requests, we inspect the URL of the requests and examine their path and parameters. For HTTP POST requests, we inspect the data stored in the request body. We report a case of ID synchronization only if an identifier is delivered to a domain different than the one that assigned it to the user originally. This analysis is performed for both first-party and third-party identifiers and in a per-website basis.

Based on empirical analysis and as described in Chapter 2.1, the majority of identifiers are stored in cookies. Thus, we parse the value of each stored cookie and look for strings that can be used as unique identifiers. If this value is a text string representing a JSON object, we get only the values stored in the key-value pairs of the object. We purposely ignore the keys found in key-value pairs since these keys rely on the API defined by the website, and do not contain any

Cookie Database		
Name	Value	Domain
tid	437f91f25a64fa45592e243aa8	trivago.co.id
...

Network Traffic
<p>General</p> <p>Request URL: https://googleads.g.doubleclick.net/pagead/viewthroughconversion/1052672223/?random=1600260088500&cv=9&fst=1600260088500&num=1&userId=437f91f25a64fa45592e243aa8&label=q46FCPnUznsQ34H69QM&guid=ON&resp=GooglemKTybQhCs0&u_h=1000&u_w=1900&u_ah=920&u_aw=1900&u_cd=24&u_his=2&u_tz=180&u_java=false&u_nplug=</p> <p>Request Method: GET</p>

Figure 3.2: Example of a first-party identifier being leaked to a third-party. A user ID assigned by *trivago.co.id* is being shared with Google.

useful information that can uniquely identify users. Treating these keys as possible identifiers would result in multiple false positives, considering that the keys would frequently appear in the parameters of GET requests. Additionally, if the object contains inner JSON objects, we recursively obtain all values in all nested levels. In addition, we see cookie values combining (with a delimiter) identifiers along with non-identifying info (e.g., timestamp, locale, etc.). Such an example is the cookie `foo={userID};15693242;en-US` that stores a user ID along with other information. We find such values appearing in less than 1% of the detected cookies, with only 0.6% of such identifiers being synced with third-parties, and therefore, choose to exclude combining values from our analysis.

To reduce false positives, we deliberately filter out values that include consent information (e.g., values of the keys `euconsent`, `eupubconsent`, `__cmpconsent` and `__cmpiab`). As described in [45], such values can be used to share a user’s consent across different CMPs or third-parties present on the page. Furthermore, we filter out frequent values that are considered common and cannot be used as identifiers. Specifically, we exclude strings that represent dates, timestamps, regions, locale, as well as, strings that end with a common file extension (e.g., `.jpg`) and strings that are URLs (e.g., start with `www.` or `http://`). We also exclude strings that have a length of 5 or less characters as they do not carry enough information to uniquely identify a user.

Finally, we filter out cookies with values which are prevalent keywords and are not used as identifiers. To construct a list of such keywords, we use a simplified puppeteer-based crawler to visit over 2.5K websites. Websites were selected randomly from the Tranco list presented in Chapter 3.1. When the crawler visits a website, it waits for a short period of time for transactions to take place and then stores both first-party and third-party cookies. We manually inspect the values of the collected cookies and identify over 80 keywords that are frequently found in cookies but cannot be used for user identification. As an illustration, this list

includes keywords such as “homepage”, “undefined”, “desktop”, “not set” and “active”, among others.

The last step in our analysis is to detect the possible identifiers in network traffic. For each string of the previous step, we examine all HTTP(S) requests targeting domains different than the one that set the cookie, and seek for an exact string match. Finding such a match would indicate that an identifier has been delivered to a different party, thus leaking the identity of the user. We search for these possible identifiers in (i) URL parameters and fragment, (ii) the body of requests and (iii) the referrer header. We tokenize the URL parameters using both default (i.e., `&`) and custom (i.e., `;`) delimiters. Finally, these components are decoded to expose values, which would be hidden due to URL encoding. For example, a cookie value `e65:4c5#d5+5b&23asd` can be used as a user identifier, but sent through an HTTP GET request, it would be encoded as `e65%3A4c5%23d5%2B5b%2623asd` and, therefore, impossible to detect.

3.4 Detecting Browser Fingerprinting

As described in [48] and illustrated in Figure 2.2, browser fingerprinting techniques, such as canvas fingerprinting, can be performed using various native methods provided by the browser’s run-time environment. For example, in the case of canvas fingerprinting, these methods are related to the HTML5 canvas element and provide the necessary functionality to draw on the canvas or collect the representation of the drawn image (e.g., `fillText`). Past work [1, 65, 20, 42] has focused on monitoring these native methods along with their returned values in order to detect browser fingerprinting. By observing the sequence of function calls along with the arguments given to these functions, one can have indications of browser fingerprinting. Additionally, searching for common arguments found in popular fingerprinting libraries can help increase the level of certainty. We argue that this method produces multiple false positives, since websites which use the native methods or HTML elements, like the canvas element, legitimately, might be marked as fingerprinting websites. Indeed, in [65] the authors explain that manual revision of results was required in order to exclude false positives.

To mitigate this, we propose a novel approach which focuses on a higher level of abstraction and does not examine the browser’s built-in (i.e., native) methods. Specifically, to detect browser fingerprinting, we perform JavaScript code profiling and search for specific function calls that indicate the presence of a fingerprinting library. This way, we successfully disregard websites that use native methods legitimately (e.g., the canvas element for web graphics). Our method reduces the number of true positives, but ensures that the results are trustworthy. Moreover, this method can be utilized by a fully-automated crawler, without the need of any manual intervention.

In particular, we analyse the open-source version of one of the most widely-used fingerprinting JavaScript libraries: FingerprintJS [32]. We extract the full

```
910 var getCanvasFp = function (options) {
911     var result = []
912     ...
913     var canvas = document.createElement('canvas')
914     canvas.width = 2000
915     canvas.height = 200
916     canvas.style.display = 'inline'
917     var ctx = canvas.getContext('2d')
918     ...
919     ...
935     ctx.fillText('Cwm fjordbank glyphs vext quiz, \ud83d\udef0', 2, 15)
```

Figure 3.3: Snippet of the popular fingerprinting library, FingerprintJS.

list of declared functions used during the process of browser fingerprinting. We then focus on functions that consist of multiple operations and require a significant number of execution cycles. This ensures that they will always be sampled by the profiler. Moreover, we ignore functions that have common names (e.g., `map` or `isIE`) and functions that can be utilized by general purpose code to perform actions not necessarily related to fingerprinting (e.g., `getRegularPlugins`). As a result, we conclude that the execution of the functions `getCanvasFp`, `getWebglFp`, `Fingerprint2` and `Fingerprint2.get` signify browser fingerprinting. These functions indicate clear intent to fingerprint the user's browser and uniquely identify them.

Next, to fully automate the detection of browser fingerprinting, we modify our puppeteer-based crawler to start with the built-in profiler tool of the Chromium browser, enabled. This is achieved using Puppeteer's ability to create a session for the Chrome DevTools protocol [5]. Additionally, we set the sampling interval of the profiler to 500 μ s, which results to 2K samples per second. The output of the Chromium profiler is a list of profile nodes. Each node contains information about samples, in addition to a unique ID and a call frame. Using this call frame, we extract the function name along with the URL of the JavaScript code that contains, and executes, the specific function. This enables us to search for fingerprinting functions, as well as identify the exact script that performs browser fingerprinting. Finally, the algorithm described in Chapter 3.3 and the browser fingerprinting detection methodology are publicly available in the released repository [56].

Chapter 4

Analysis of Consent

In this section, we present our measurements and analyze the behavior of websites across three types of visits: when consent is (i) rejected (**Reject All**), (ii) granted (**Accept All**), and (iii) not responded to (**No Action**). We study how websites change their user tracking behavior depending on the consent provided (or not) via the number of third parties they integrate with, as well as, (Chapter 4.1), first-party ID leaking (Chapter 4.2), third-party ID synchronization (Chapter 4.3) and browser fingerprinting (Chapter 4.4) performed. We also study how other factors such as a website’s popularity (Chapter 4.5) and the hosting country (Chapter 4.6) may affect the intensity of the website’s tracking behavior.

4.1 Consent and Third-Parties

First, we study how websites change their user tracking behavior depending on the consent action that the user performed. To that extent, we measure the number of unique third-parties interacting with the first-party through application-level network traffic. In Figure 4.1, we plot the distribution (min, 25th percentile, median, 75th percentile, max) for the three different consent actions. We observe that in the **No Action** and in the **Accept All** case, there are 16 and 19 third-parties interacting with the median website, respectively. Surprisingly, we observe that in the **Reject All** case, there are more (i.e., 17) third-parties running in the median website than when the user performs no action at all and that it may reach up to 29 distinct third-parties for the 75th percentile. This suggests that simply interacting with the consent manager has an impact on the number of third-parties in the visited website.

To verify our hypothesis, we perform two sided non-parametric Kolmogorov Smirnov tests, which compare the underlying distributions of two samples by quantifying the distance between their empirical distribution functions. We find that the KS statistic value is $D_{(noaction,rejectall)}=0.038$, $D_{(rejectall,acceptall)}=0.061$ and $D_{(acceptall,noaction)}=0.097$ for the three pairs of consent actions. In all comparisons, the p-value was less than 10^{-10} indicating that the three distributions are indeed

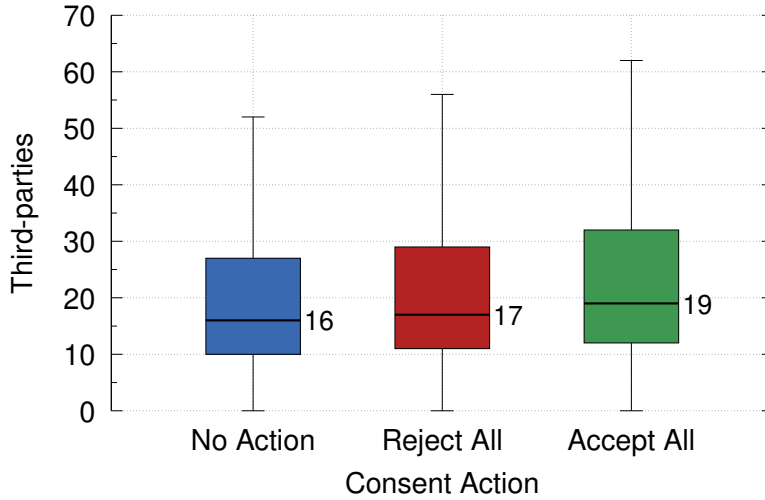


Figure 4.1: Number of third-parties running on the website during the three different types of visits (i.e., different consent actions). Surprisingly, for the median website, in the `Reject All` case there were more (i.e., 17) third-parties running than in the `No Action` case (i.e., 16).

statistically different. This verifies that the behavior of websites changes based on the consent action and that there are more third-parties interacting with the median website when consent was denied than when the user did not interact with the consent banner at all.

Additionally, we measure the most popular trackers that websites interact with. We classify domains as trackers based on the Tracker Protection Lists [31] provided by *Disconnect*, a popular technology company that focuses on transparency and control over personal information on the Web. Surprisingly, we find that 1,548, 1,562 and 1,718 websites interact with trackers in the `No Action`, `Reject All` and `Accept All` cases, respectively. As expected, more websites interact with trackers when they have received valid consent from the user. This behavior is expected and inline with the regulations. Surprisingly, there are slightly more websites that issue requests towards third-party trackers when the user denied consent than when the user did not perform any action at all. This finding, not only highlights that simply interacting with a consent manager in a website increases the number of third-parties that will interact with the website, but also that some additional tracking services will be involved in this tracking process.

4.2 Sharing User IDs with Third-Parties

In our next experiment, we set out to explore the cases where a first-party identifier (e.g., cookie, device ID [51]), previously set by the visited website, is being leaked to various third-parties. Therefore, we measure how many first-party ID leaking

Table 4.1: Number of websites detected (i) leaking their first-party user identifiers and (ii) having third-parties that perform synchronizations of user IDs.

Consent Action	Websites engaging in first-party ID leaking	Websites with third-party ID synchronization
No Action	14,238 (52.38%)	6,533 (24.03%)
Reject All	15,334 (56.41%)	7,123 (26.20%)
Accept All	17,764 (65.35%)	8,048 (29.61%)

operations are being performed in a website as a function of the three aforementioned user choices. Table 4.1 summarizes our findings. One would expect that no such operations take place before the user makes a choice (i.e., **No Action** case), as well as when the users denied consent (i.e., **Reject All**). However, among the websites that present their users with a consent banner, we found over 14,000 of them performing first-party ID leaking even before their users had the opportunity to register their preferences (**No Action** case). To our surprise, when users **Reject All** cookies, the first-party ID leaking only gets worse, with more than 56% of websites engaging in it.

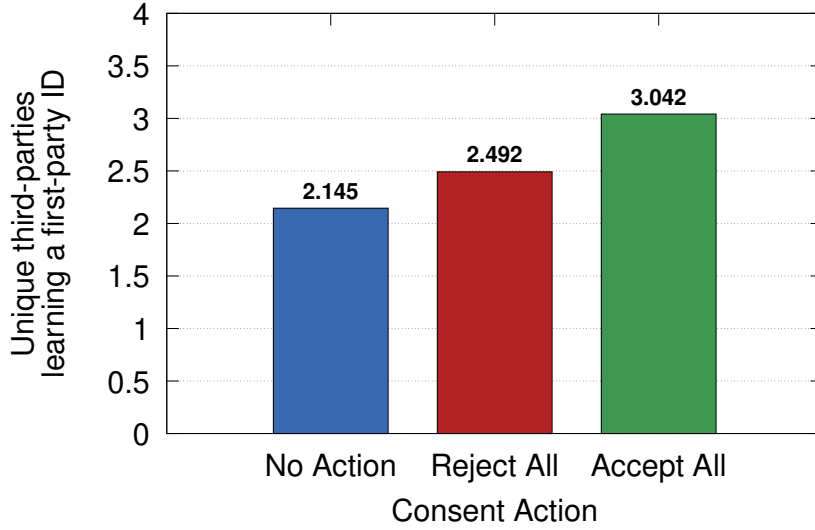


Figure 4.2: Average number of unique third-parties learning a user identifier. A user’s browser leaks first-party identifiers to 2.14 third-parties, on average, even before the user accepted or rejected consent.

Next, we explore the extent of these leaks. Therefore, we study these websites that perform first-party ID leaking in at least one of the three visits and report the average number of third-parties learning a user identifier. Figure 4.2 illustrates our findings. Specifically, we find that, on average, when a user visits these websites there are 2.14 third-parties that learn a user identifier even before the user has the

Table 4.2: Top 5 third-parties that learn the highest number of first-party identifiers per consent action in our dataset. For each party we compute the percentage of leaked first-party identifiers that they learn.

#	No Action	Reject All	Accept All
1.	facebook.com 18.87%	facebook.com 18.29%	facebook.com 19.48%
2.	google-analytics.com 18.85%	google-analytics.com 17.28%	google-analytics.com 15.99%
3.	bing.com 9.64%	bing.com 8.84%	bing.com 10.27%
4.	hubspot.com 6.66%	doubleclick.net 6.60%	doubleclick.net 6.82%
5.	doubleclick.net 4.68%	hubspot.com 5.86%	hubspot.com 5.99%

opportunity to make a decision. To make matters worse, if the user chooses to deny consent and reject all cookies, first-party identifiers may be leaked to even more third parties (on average 2.49). We observe that the difference in third-parties that learn a first-party identifier in the **Reject All** case and in the **Accept All** case is rather small. Precisely, in the average website, choosing to **Accept All** cookies leaks first-party identifiers to 3.04 distinct third-parties while choosing to deny consent (i.e., **Reject All**) leaks identifiers to 2.49 third-parties. This difference is hardly significant with only 25% less third-parties, suggesting that more than 75% of the third-parties that will learn a first-party identifier, do so without the user’s consent.

Finally, in Table 4.2, we show the top 5 third-parties in our dataset that learn the most first-party identifiers across all websites for each of the three consent options. For this experiment, we perform our analysis by extracting the base domain (i.e., eTLD + 1) of each third-party learning an identifier. As expected, we find that these parties are popular Web trackers, thus supporting our finding of sophisticated tracking taking place even when the user denies consent. Facebook with its social plugin, Google with its analytics tracker and ad-exchange (DoubleClick) modules, and Microsoft (Bing) occupy the top positions in all three consent options. Interestingly, we find that Facebook learns approximately 19% of the leaked identifiers regardless of the consent action. In addition to this, we find all of these third-parties in Disconnect’s Tracker Protection Lists [31].

Before proceeding with more experiments and results, let us take a step back and reflect on “what is the point of the consent provided by the end users?”. It is true that all studied websites asked the user’s consent and even offered some choices. Thus, from a GDPR-compliance point of view, these websites asked for the user’s consent in an attempt to establish a legal basis. However, before the user had any opportunity to respond and provide their preference, these websites started leaking identifiers. To make matters worse, if the user chooses to reject all data processing purposes than to take no action at all, the first-party ID leaking

Table 4.3: Top 5 third-parties with highest number of third-party synchronizations per consent action in our dataset.

#	No Action	Reject All	Accept All
1.	doubleclick.net 21.15%	doubleclick.net 21.47%	doubleclick.net 20.22%
2.	everesttech.net 13.21%	everesttech.net 12.10%	everesttech.net 10.89%
3.	scorecardresearch.com 10.59%	facebook.com 9.95%	facebook.com 9.61%
4.	facebook.com 10.15%	scorecardresearch.com 9.61%	ad.gt 9.54%
5.	taboola.com 9.68%	google-analytics.com 8.30%	taboola.com 8.49%

only intensifies. In fact, more than 75% of first-party identifier leaks happen despite the fact that users have chosen to deny consent. Clearly, these aggressive first-party ID leaking operations are not compatible with what users expect when they are asked to give their consent.

4.3 Third-party ID synchronization

As described in Chapter 2.2, apart from sharing the first-party identifiers they assign to the visiting users, websites may also host third-parties which synchronize the different user IDs they use for the same users. As shown in Table 4.1, from the websites that present their users with a consent banner, we found over 6,500 websites of them hosting third-parties that conduct synchronization of identifiers before users had the opportunity to register their choices (**No Action** case). If users prefer to deny consent, then even more websites (26%) engage in third-party ID synchronization. Although consistent with the finding of the previous section (i.e., first-party ID leaking), this fact highlights and supports our finding that websites employ sophisticated forms of tracking totally disregarding user consent preferences.

To quantify the extent of this phenomenon as a function of the three consent actions, we measure the average number of unique third-parties learning a user identifier assigned by a different third-party. When the user takes **No Action**, the browser engages in 3.51 synchronizations, on average. This means that when the user is asked for GDPR compliance, and before even responding, their browser already leaked at least one identifier assigned by a third-party to more than three other third-parties. To make matters worse, if the user responds negatively and chooses to **Reject All** cookies, their cookies may get synced with even more third-parties (3.91 on average). Finally, the third-party synchronization operations increase rapidly to 4.86 in the **Accept All** case.

Similar to Chapter 4.2, we present in table 4.3 the top 5 third-parties conducting the highest number of synchronizations across websites, for each of the consent options. Again, we find that popular tracking domains are most likely to engage in third-party ID synchronization. Google’s ad-exchange platform `doubleclick.net` and Adobe’s tracker `everesttech.net` are the top two third-parties in all three consent actions. Next, we group third-parties based on the company that operates them according to information found in the Tracker Protection Lists [31]. We find that in all three consent options, Google learns over 40% of the synchronized third-party identifiers. Consequently, Google is the dominant player in the cookie synchronization ecosystem participating in almost half of the synchronization operations. By performing aggressive and sophisticated tracking, Google is able to track users across numerous websites and build a very accurate profile of users on the Web.

In total, we find that websites with embedded third-parties that synchronize the identifiers they have assigned to the same user, force browsers to engage in 3.51 synchronizations, on average, even before the users had any chance to accept or reject consent. This way, these third-parties can later target specific groups of users [2], sell their data [12], or use these data in ad-auctions [60, 54]. We argue that this type of leakage is worse than first-party ID leaking, since (i) it is not in the immediate control of the websites themselves and (ii) via this mechanism, third-parties that are not present on the website can be alerted of a user’s presence and online behavior.

4.4 Browser Fingerprinting

In our next experiment, we set out to explore whether websites track users differently using browser fingerprinting, given the different user responses to the consent banners. By following the detection methodology described in Chapter 3.4, we detect websites performing browser fingerprinting across the different types of visit. Table 4.4 presents our findings and, as we can see, the action of the user has no significant impact on the websites’ fingerprinting operations. Specifically, 279 websites perform browser fingerprinting even before the user had the opportunity to respond to the consent request. Even worse, if the user selects to **Reject All** cookies, even more websites (285) engage in browser fingerprinting. This number is increased to 330 when the user chooses to **Accept All** cookies.

It is interesting that the difference between the **Reject All** and **Accept All** cases is only 15%. Moreover, we see 73.5% of the fingerprinting websites perform browser fingerprinting no matter what the user consent action is. This implies that for websites that track users using browser fingerprinting, user choice makes very little to no difference. This is evident in the fact that only 2% of these websites wait for user’s action before starting their fingerprinting operation.

Furthermore, we find that only 13.9% of these websites comply with legislation and perform browser fingerprinting only when the user gives consent. On the

Table 4.4: Websites performing browser fingerprinting.

Description	Volume	% of total
No Action	279	1.03%
Reject All	285	1.05%
Accept All	330	1.21%
In at least one consent action	336	1.24%
In all 3 cases	247	73.5%
Only in Accept All case	47	13.9%
Only in Reject All case	3	0.9%
Wait for action	7	2%

other hand, we find 3 websites that perform browser fingerprinting only if the user denies consent and reject all data processing purposes. It is apparent that these websites are using browser fingerprinting as a fallback mechanism in the case that they are not allowed to set a cookie on the user side. However, it is important to notice that based on Article 4/Recital 30 [15] and explained in Chapter 2.4, GDPR regards the process of any user identifying information and is not limited to cookies. Consequently, these websites violate the legislation that applies to citizens of the European Union. Altogether, we find that even though websites ask users for consent, they do not take the response into account when performing browser fingerprinting.

4.5 Does website popularity matter?

In our next experiment, we explore whether a website’s popularity impacts how the website handles the user’s choices. For this reason, we cluster websites into buckets based on their popularity according to the Tranco list. The first bucket contains the 50K most popular websites in the Tranco list, the next bucket contains the next 50K (i.e., ranked 50K-100K), etc. In this experiment, we study 17,900 websites that performed first-party ID leaking in any of the three visits. In Figure 4.3, we show the extent of first-party ID leaking for the different buckets for the three cases we study: **No Action** (blue bar), **Reject All** (red bar), and **Accept All** (green bar). For illustration purposes, we plot only the first ten buckets. We see that as the popularity of the website decreases (right part of the plot), all bars tend to decrease, implying that the magnitude of tracking through first-party ID leaking decreases as well.

To verify this hypothesis, we normalize the values so that the **Accept All** case corresponds to the maximum tracking intensity (i.e., 100%). Figure 4.4 illustrates our findings. In this case, we observe that for the **No Action** (i.e., blue bars) and **Reject All** (i.e., red bars) cases, there is a slightly increasing trend to the right. That is, less popular websites tend to be slightly more aggressive in disregarding

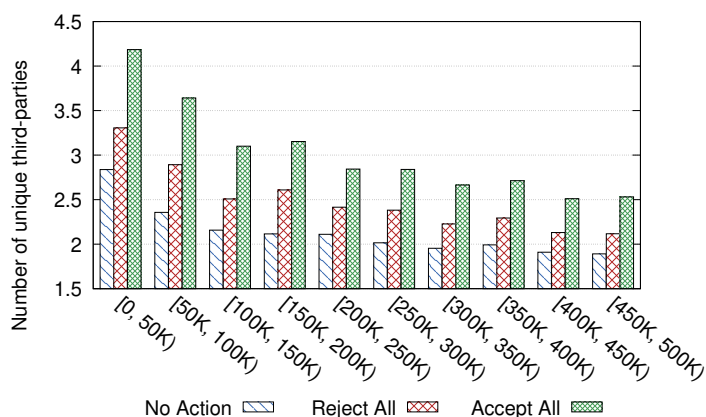


Figure 4.3: Mean number of third-parties involved in first-party ID leaking per website rank.

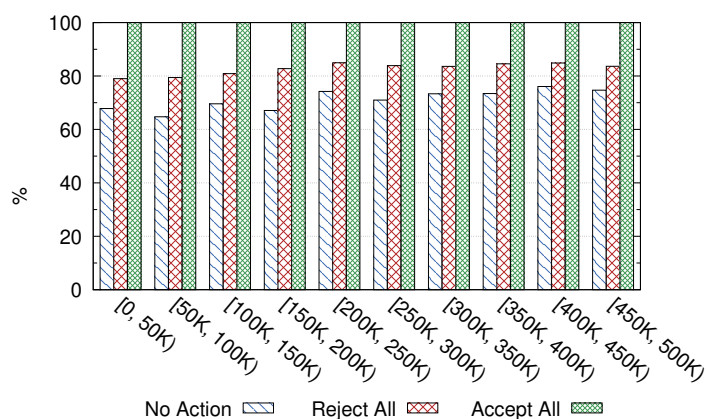


Figure 4.4: Unique third-parties in ID leaking per website rank (normalized).

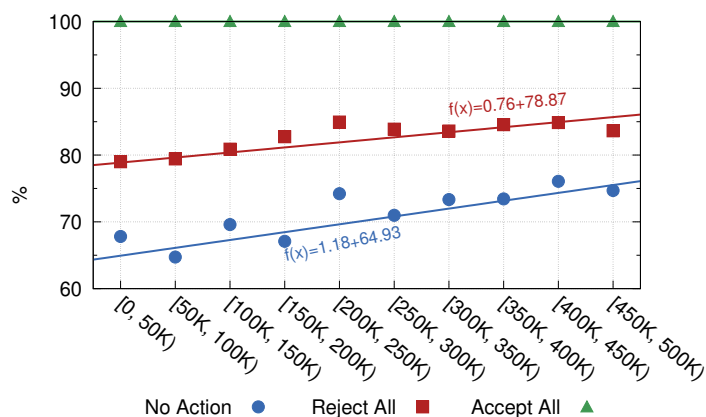


Figure 4.5: Unique third-parties in ID leaking per website rank (linear regression).

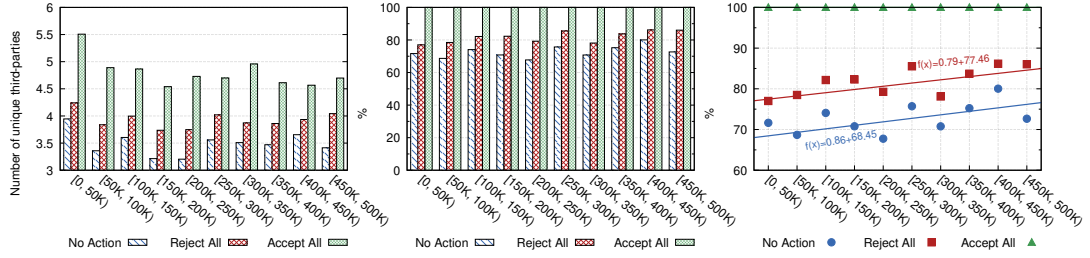


Figure 4.6: Unique third-parties involved in ID synchronization per website rank. Figure 4.7: Unique third-parties involved in ID synchronization per website rank (normalized). Figure 4.8: Unique third-parties involved in ID synchronization per website rank (linear regression).

Figure 4.9: third-party ID synchronization as a function of the website’s popularity. (a) Average number of unique third-parties involved in synchronizations per rank range of the website. (b) This figure plots the same information as Fig. 4.6, with the exception that all **Accept All** values are normalized to 100%. (c) In this figure exception that **Reject All** and **No Action** points have been fitted with a straight line. The line suggests an increasing trend implying that less popular sites are more aggressive at disregarding user choices.

user choices. For example, the 50K most popular websites do 67% of their first-party ID leaking before the user makes any choice, while for less popular websites in the range [400K, 450K), this tracking behavior increases to 76%. In Figure 4.5, we highlight this clear trend by showing an interpolation of the results. This figure plots the same information as Fig. 4.4, with the exception that **Reject All** and **No Action** points have been fitted with a straight line. The line suggests an increasing trend implying less popular sites are more aggressive at disregarding user choices. In both consent actions (**No Action** and **Reject All**), we observe a positive slope. The coefficient of determination is $R^2=0.42$ and $R^2=0.04$ for the two actions respectively.

Similarly, we perform the same experiment for 8,301 websites that were involved in third-party ID synchronization operations for any of the three consent options. In Figures 4.9, we see the same trend across the popularity buckets of websites hosting synchronizing third-parties. Therefore, we conclude that even though more popular websites perform more aggressive tracking and leak more identifiers, it is evident that less popular websites tend to be more aggressive at disregarding users’ consent choices and engage in both first-party ID leaking and third-party ID synchronization.

4.6 Does the hosting country matter?

Next, we study how the websites hosted in different countries treat user consent. As a result, we study the country code top-level domain (ccTLD) of each website we visited. We present the results for first-party ID leaking in Figure 4.10. For this experiment, we focus on websites that perform first-party ID leaking in at least one of the three consent actions. Additionally, we ignore ccTLDs with a small number of websites, since the small populations will not result in statistically significant results. As an illustration, there are only 9 websites from Luxemburg (i.e., *.lu* ccTLD) that perform first-party ID leaking in any of the three visits. On the contrary, we find over 440 German websites (i.e., *.de* ccTLD) leaking first-party identifiers in our dataset.

We observe that Europe-based websites are more likely to respect the choices of the users and not track them aggressively. For example, we discover that websites from France (fr), Denmark (dk), the Netherlands (nl), Austria (at) and Germany (de) leak first-party identifiers to less third-parties than websites with non Europe-based ccTLDs like Canada (ca) and Australia (au).

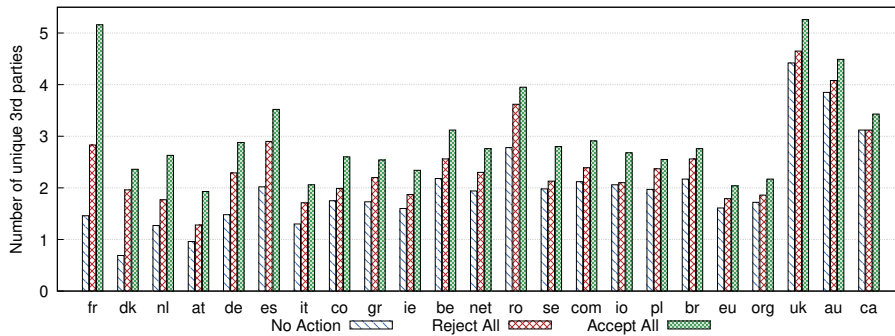


Figure 4.10: Number of unique third-parties learning a first-party identifier as a function of the top-level domain per country code.

In Figure 4.11, we normalize the results based on the **Accept All** case. This enables to compare websites and ccTLDs with different levels of leaks to third-parties. Websites in the right part of the figure have similar tracking behavior when it comes to first-party ID leaking in the three consent options. This indicates that these websites tend to disrespect user choices. For example, the difference between **Reject All** and **Accept All** in websites from Australia (au) and Canada (ca) seems to be negligible. As a result, the choice that the user makes when interacting with consent banners in such websites makes little to no difference. Surprisingly, we see that the ccTLD of *.eu* is on the right side of the figure. This strongly suggests that there is an increased number of websites with this specific ccTLD which are not yet fully compliant with GDPR.

On the other end of the spectrum, websites on the left part of the figure seem to respect user choices more. For example, the difference between **Reject All** and

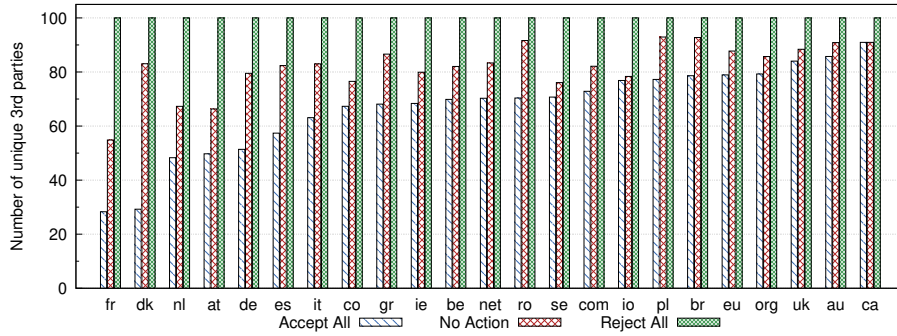


Figure 4.11: Normalized number of unique third-parties learning a first-party identifier. This figure plots the same information as Figure 4.10, with the difference that the max value (**Accept All**) is normalized to 100%. This enables us to compare websites that have different magnitudes of leakage.

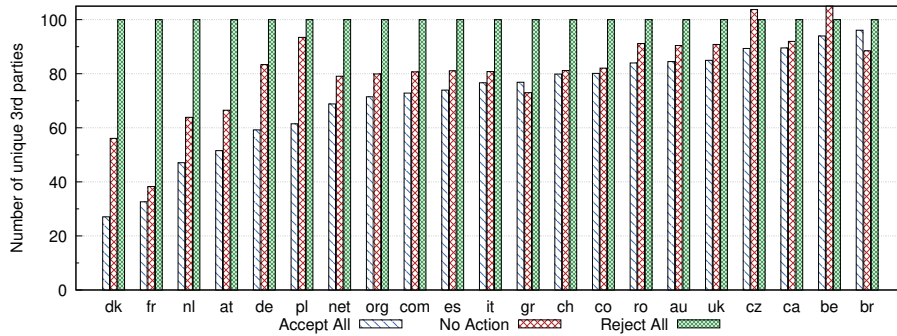


Figure 4.12: Normalized number of unique third-parties engaged in third-party ID synchronization.

Accept All for French websites (i.e., *.fr* ccTLD) is close to 50% and between **No Action** and **Accept All** is more than 70%. This is a clear indication that such websites take the user’s preference regarding data processing into consideration and display extensively different behavior based on the user’s choice in the consent banner. Thus, although not perfect, user choices for the websites on the left part of the figure have a meaningful effect, in contrast to websites on the right part.

We find similar results for third-party ID synchronization and report our findings in Figure 4.12. In this experiment, we focus only on websites that perform third-party ID synchronization in any visit. Again, we see that European websites hosted in countries like Denmark (dk), France (fr), the Netherlands (nl) and Austria (at) tend to perform less third-party ID synchronizations when there is no explicit consent given by the user (i.e., **No Action** and **Reject All** cases). On the other hand, we find that for websites with non Europe-based ccTLDs like *uk*, *ca* and *br*, user choice has a much smaller impact. Surprisingly, we see that two

European ccTLDs (*cz* for Czech Republic and *be* for Belgium) are on the right part of the figure, indicating that they do not respect user choices as much as other European websites. To make matters worse, websites with these ccTLDs perform more synchronizations when users deny giving consent. One possible explanation is that these websites use third-party ID synchronization as a fallback tracking mechanism when the user denies consent, as described for browser fingerprinting in Chapter 4.4. Additionally, specific behaviors of multiple websites in one country can be attributed to the implementation and practices of popular advertising networks and tracking services in these countries.

Finally, we investigate the top-level domains of fingerprinting websites. We examine the 336 websites that perform browser fingerprinting in any of the three consent actions. We discover that 79% of these websites use country-independent top-level domains with the great majority (72%) of them using the *.com* TLD. The second most popular top-level domain we discovered was *.net*. Regarding countries and ccTLDs, we find that Germany is the country with the most fingerprinting websites (4%).

Altogether, we believe that our analysis of tracking mechanisms per country code reveals significant discrepancies across the European Union. These results highlight the lack of effort from specific local governments or agencies regarding the digital privacy rights of their citizens. Additionally, our analysis emphasizes that there is a long way to go before GDPR is actually enforced, especially in a global scale. Particularly, GDPR was adopted in 2016 and came into force in 2018. Even though website administrators and publishers had enough time to prepare before it came into effect and even more time to comply with the regulation after it was implemented, we find that there are still numerous websites that do not respect user consent.

Chapter 5

Ineffective Consent: Edge cases

To further understand the Web ecosystem, as well as the tracking behavior displayed by websites, we selectively present some extreme cases of ineffective consent, meaning that websites do not take the user’s consent into account and engage in aggressive tracking. Before reporting the following edge cases, we re-crawled these websites and manually evaluated the results in order to verify our findings.

In our dataset, we observe 73 websites that interact with over 100 unique third-parties each, in at least one of the three types of visit. One such example with extreme behavior is *laprovence.com*. When a user visits the website and grants consent for all data processing purposes (i.e., **Accept All**), the website interacts with 159 different third-parties and performs synchronization for multiple identifiers with 59 of these parties. We observed the values of 37 unique third-party cookies being leaked to third-parties different from the cookie’s owner. In the **Reject All** case, the website interacts with 80 third-parties, and performs synchronization for at least one identifier with 16 of them. Interestingly, when the user lands on the website with a clean session and performs **No Action**, but simply waits for a big period of time, the website interacts with 97 third-parties and performs synchronization with 29 of them. This is clear indication that there is indeed a total disregard of user consent observed, even in popular websites.

Regarding first-party ID leaking, we find that multiple websites store a cookie labeled as “necessary”, but then proceed to leak its value to various third-parties. For example, *harryanddavid.com* leaks the values of 28 different first-party cookies in the **Reject All** and **No Action** cases. We investigate the exact cookies being leaked and verify that they are indeed identifiers assigned to the user with names such as `user_id` and `sessionID`. Furthermore, *diariodepontevendra.es* and *asivaespana.com*, in the **Reject All** and **Accept All** cases, respectively, perform ID leaking with 38 different third-parties for more than one identifier.

In addition to this, *camer.be* interacts with 91 unique third-parties in the **No Action** case, 94 in the **Accept All** case, and surprisingly with 131 in the **Reject All** case. For the **Reject All** case, this website is also involved in a major third-party ID synchronization operation. At the time of crawling, the website interacted

with the third-party `taboola.com`, a popular advertising company, which stored a cookie with name `t_gid` and value `884d05cc-335c-4226-ab94-7ab6114fef6a-tuct65bfbc8`. This value, which follows the UUID format [43] and can uniquely identify users and devices, was sent to 20 other third-parties. One very interesting finding is that this cookie is stored only when the user declines consent (i.e., `Reject All` case). Additionally, `taboola.com` is reported as one of the top third-parties involved in third-party ID synchronization in Table 4.3.

Similarly, `cnnturk.com` is also involved in a major third-party ID synchronization operation. Specifically, when the user lands on the website with clean state, a third-party called `lijit.com` stores the cookie `_ljtrtb_42`. The value of this cookie is then sent to 21 other third-parties. Interestingly, this behavior is observed only after the user has interacted with the consent form (i.e., `Reject All` and `Accept All` cases). One example value of this cookie that we observed during the `Accept All` case is `c98d9202-8774-4e11-8c90-99d9cb879930-tuct65c0de5`, which can be used to uniquely identify a user. Note that `lijit.com` is an ad-serving domain, which can be found in multiple blacklists for tracking domains.

Finally, `glamour.com` leaks a unique identifier which is set as the value of a first-party cookie. Specifically, when a user lands on the website, a cookie called `CN_xid` is stored, with one example value being `73a4ff1f-ff45-4943-bdaa-73658b00bd42`. Then, this value is sent to exactly 21 unique third-parties. The third-parties that receive the value of the cookie are exactly the same for all 3 types of visits. An interesting finding is that the third-parties that receive this value are not only domains known for advertising and analytics (e.g., Google Analytics and Doubleclick), but also legitimate and mainstream content-serving websites like `vogue.com` and `wired.com`.

Chapter 6

Related Work

The recent increased interest of regulators and governments around the privacy rights of Internet users did not result only in legislation like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), but also in an important body of research. Tracking mechanisms such as HTTP cookies [39, 8, 21] and browser fingerprinting [17, 48, 41, 76] have been the prime focus of academic work for a long period of time. With the advent of privacy laws, academic work focused on how identifiable information stored in cookies is maintained and does not examine the compliance of deployed stateless tracking (e.g., browser fingerprinting).

In [27], the authors investigated the legal compliance of purposes for 20K collected third-party cookies. Their findings show that purposes declared in cookie policies do not comply with the purpose specification principle in 95% of cases. In [13], the authors collected cookies from the Alexa Top 100K websites and compared their cookie behavior from different vantage points, to investigate whether there are differences in cookie setting when accessing Internet services from different jurisdictions. Additionally, they studied whether cookie setting behavior has changed over time by comparing today's results with a dataset from 2016.

In [14], the authors studied popular websites of all member states of the European Union and investigated changes in the privacy policy presented to users before and after the GDPR came into effect. Their finding was that the GDPR brought a shift on the Web with websites becoming more transparent but there is still a lot of work to be done for users to be able to properly deny consent. Hils et al. performed a longitudinal analysis in [29] in order to study the rise of Consent Management Providers. They found that privacy laws have urged websites to adopt the services provided by CMPs, and that CMPs are becoming extremely popular with a vast growth, especially in popular websites.

In [67], the authors performed an evaluation of the tracking activities performed in 2K high-traffic websites, hosted both inside and outside the EU. Specifically, they evaluated the information presented to users and the actual tracking implemented through cookies. Their results show that the GDPR has impacted website

behavior in a truly global way. US-based websites behave similarly to EU-based ones, while third-party opt-out services reduce the amount of tracking, even for websites which do not put any effort in respecting the GDPR. On the other hand, they show that cookies can identify users when visiting more than 90% of the websites they crawled, and they encountered a large number of websites that present deceiving information, making it very difficult, if at all possible, for users to avoid being tracked. Similar to this work, in [18], the authors crawled 1.5K EU, US, and Canadian websites from 18 countries and analyzed the cookie notices they encountered. Using a series of regression models, they found that a website's Top Level Domain explains a substantial portion of the variance in cookie notice metrics, but the users vantage point does not, which means that websites follow one set of privacy rules for all their users.

In [47], the author performed an analysis of the privacy notices presented in different platforms, specifically personal computer browsers, mobile browsers and mobile applications. Using a privacy-oriented browser for personal computers and a privacy monitoring app for mobiles, they detected tracking activities in the landing page of the 150 most popular EU websites. Similar to our results, their finding was that tracking activities take place before the user has the chance to interact with the banner. Moreover, they found that the number of tracking activities is similar in personal computers and mobile devices. In [74] authors studied the impact of the legislation on cookie syncing between third-parties. They show that the general structure of how the entities are arranged is not affected by the GDPR, but the new regulation has a statistically significant impact on the number of connections that shrunk by 40% in the GDPR era.

In an effort closest to ours, Matte et al. analyzed the GDPR and ePrivacy Directive across 23K European websites to identify legal violations in implementations of cookie banners based on the storage of consent [45]. That is, they (i) capture the user's choice (consent or not), (ii) measure whether the websites register the same response as the user's choice, (iii) measure whether websites register any response *before* the users click their preference. They found that 141 websites register positive consent even if the user has not made their choice; 236 websites nudge the users towards accepting consent by pre-selecting options; and 27 websites store a positive consent even if the user has explicitly opted out. Performing extensive tests on 560 websites, they found at least one violation in 54% of them. Although our work and [45] share similar goals, they clearly have significant differences. First, although [45] focuses on cookies as the main tracking mechanism, in this work, we focus on post-cookie tracking mechanisms including browser fingerprinting, first-party ID leaking and third-party ID synchronization. In this aspect, we explore whether sites use such tracking mechanisms to bypass any consent the user has provided for cookies. Second, [45] focuses on whether the Consent Management Platforms registers the same response as the user's input. We follow a different methodology and measure *not* the response registered, but the actual tracking mechanisms that are activated when the users access a website.

Finally, there has been a lot of work that focuses on the design of consent

banners. In [75], authors studied the common properties of the graphical user interface of consent notices and conducted three experiments with more than 80K unique users on a German website, to investigate the influence of notice position, type of choice, and content framing on consent. Their results show that (i) users are more likely to interact with a notice shown in the lower left part of the screen, (ii) users are willing to accept tracking compared to mechanisms that require them to allow cookie use for each category or company individually, (iii) the widespread practice of nudging has a large effect on the choices users make. Similarly, Nouwens et al. studied in [52] the effect that the design of consent banners has on the final choice that users make regarding data processing purposes. Their analysis was performed on the top 10K most popular websites in the UK and they found that only 11.8% of the evaluated websites satisfy the requirements set by legislation. In [69], the authors focus on popular news outlets and manually analyzed 300 consent banners. They detected that there are still dark patterns, such as interface interference and forced action, that attempt to deceive users.

Chapter 7

Discussion

7.1 GDPR compliance

One question that comes to mind is whether the websites discussed in Chapter 4 are in violation of the GDPR and the ePrivacy Directive. Obviously, one cannot make such a general statement for all the websites studied in this thesis. Such violations should be studied on a case-by-case basis. Even further, each website is different, and may have a legal basis to collect user data that goes beyond the user consent. What we identify in this paper is a *disparity* between (i) what the users perceive about the collection of their data, and (ii) what some websites implement with respect to data processing. Indeed, by being shown a consent banner, users perceive that they are being asked to give their permission to the website to collect and process their data. Even further, when they are given several choices, users feel that they are empowered to give a fine-grain permission, which will obviously be taken into account.

Unfortunately, this perception of the users is completely different from what various websites implement. In this thesis, we discovered that several websites collect (and share with third-parties) information about their users, even before the users had the opportunity to register their preference. Even worse, when users stated that they would like to reject all cookies, collection of their data intensified. Indeed, each website is ultimately responsible for the consent asked from their visitor. However, it is not obvious if the legal responsibility is shifted to the Consent Management Platform (CMP). Nonetheless, and considering our results, it is hard to believe that all these publishers do not respect the users' consent choices without intention (e.g., due to software bug, bad developer practices or wrong integration with their CMP).

Interestingly, existing literature, websites and blog-posts around the GDPR and changes it brought on the Internet and user tracking [71], focus solely on how identifiable information stored in cookies is maintained. However, as highlighted here, the GDPR is not only about cookies. Instead, we aim to increase user awareness regarding the GDPR (non)compliance of deployed stateless (i.e., cookie-less)

tracking, and influence a change in language used in consent request statements to be GDPR-compliant and reflect closely what the websites do in reality, in comparison to what is explained to the user.

Furthermore, our analysis of tracking per country code reveals significant discrepancies across EU (or not) countries. These results highlight the lack of effort from specific local governments regarding the digital privacy rights of their citizens. Our results can motivate them to take action and increase the GDPR enforcement in order to make websites hosted in their countries aligned with the rest of EU countries, with respect to the GDPR compliance.

7.2 Outbound Information

Although user tracking without user consent is generally undesirable, in this paper, we studied some sophisticated approaches to user tracking (such as first-party ID leaking and third-party ID synchronization) which involved not only data collection, but also data sharing with third-parties. Indeed, both approaches, provide to third-parties identifiers associated with the current user. In this way, third-parties will be able to know that this user has visited the specific website (even if they are not embedded in that website). To make matters worse, this happens even before the user has given any permission for data collection on the cookie consent banner. To put it simply: the website has already informed third-parties that this user has just visited, while the user still makes up their mind whether to give consent for data collection or not. Thus, the user is asked for consent to something that has already happened and it will keep on happening even if the user denies consent.

7.3 Edge Cases

Someone could argue that the edge-cases studied in this thesis are momentary, and cannot be held against websites as proof of non-GDPR compliance. In fact, according to [73], the third-parties in a website might change across different visits. However, even though we acknowledge the dynamicity of websites, we made a best effort to provide results that were repeatable across multiple crawls. In fact, changes in third-parties embedded in a website could change their intensity of tracking. We anticipate such changes are transient and infrequent in websites, and that high intensity of tracking is repeatable.

7.4 Methodology

The methodology we presented in this paper can be transformed into an auditing tool for regulators, stakeholders and privacy-policy makers, for verifying compliance with the GDPR, CCPA, ePrivacy Directive, and users' privacy rights. Our approach links together the (i) requested user consent with (ii) actions taken by the website based on the particular consent given. Apart from these entities, browser

vendors have already shown interest in blocking bad policies on websites [10, 77, 36] and our methodology can help towards exactly these goals. Specifically, by following our methodology, browser vendors can detect at run-time stateless device fingerprinting attempts and compare these actions with the given user consent.

7.5 Limitations

We made considerable effort to develop a browser fingerprinting detection methodology which can be deployed as a fully automated component and without the need of any manual intervention. Nevertheless, we understand and acknowledge that this approach will miss any scripts which have been minified or obfuscated, since in their nature, these processes change the declared function names. Finally, our methodology is able to yield more accurate and credible results in the expense of the number of true positives. Since the scripts detected by our methodology have dire changes in the results and findings of this thesis, we make this sacrifice willingly for greater confidence in our findings.

A careful reader might find the consent banner detection rate very low. Indeed, for European users, it seems that every visited website contains such a banner. Our methodology depends on the Consent-O-Matic browser extension, which has been built to only detect consent banners of popular Consent Management Providers. Consequently, our analysis excludes custom banners deployed by a vast number of websites. Additionally, the reproducibility of our work depends on the development of this extension. As websites, legislation and technologies evolve, so do the consent banners of CMPs. If the extension does not follow up with the changes, it won't be able to detect and properly handle consent banners.

Finally, we acknowledge the fact that our methodology is limited only to the landing page of the evaluated websites. Prior work has proven that crawling websites more deeply can have a drastic increase (about 36%) in the number of used cookies [73] and that prior studies had used suboptimal crawling methods [4] since they did not examine internal pages of visited websites. We understand that while crawling the internal pages of websites, even more advanced tracking mechanisms might be observed. Nonetheless, we refrain from doing so and visit each website as less as possible in order to ensure that we do not affect its performance or offered service. We further discuss ethical aspects regarding our crawling process in Chapter 9.

Chapter 8

Conclusion

Over the past couple of years, an increasing number of websites have started to present users with consent banners, i.e., pop-up frames that ask for the user’s permission to process personal data and store cookies in the browser. Such banners provide a variety of choices including (i) accept all, (ii) reject all, and (iii) accept some cookies. In this paper, we study whether the websites that employ such consent banners, actually track their users using “non cookie” approaches including first-party ID leaking, third-party ID synchronization and browser fingerprinting. Our results indicate that, indeed, websites that contain such banners, aggressively track their users even without their consent.

In our experiments, we found over 15,00 websites that track their users using first-party ID leaking. Even further, this tracking happened despite the fact that users of these websites had rejected all cookies and denied consent in the banner. In fact, most of these websites had started the first-party ID leaking tracking even before the users had any opportunity to interact with the consent form and register their choice.

Therefore, we highlight a significant gap between what users expect to happen when they see a consent banner in a website, and what several websites do as a result of the user’s choices. We feel that research like this helps increase transparency on the Web and expose websites which do not correspond to users’ expectations and do not comply with legislation.

Future work could focus on even harder questions such as: How should third-parties connect into CMP prompts? Is it intentional that some third-parties only take action on “reject all” option? If yes, why? Are some CMPs better than others with respect to GDPR compliance? Are all these privacy violations the website’s, the CMP’s, or the third-party’s fault?

Chapter 9

Ethical Considerations

The execution of this work has followed the principles and guidelines of how to perform ethical information research and the use of shared measurement data [38, 66]. In particular, this study paid attention to the following dimensions.

We keep our crawling to a minimum to ensure that we do not slow down or deteriorate the performance of any web service in any way. Therefore, we crawl only the landing page of each website and visit it only the minimum required times. Additionally, we do not visit websites that state they do not allow bots and crawlers in their robots.txt files. We do not interact with any component in the visited website, and only passively observe network traffic. In addition to this, our crawler has been implemented to wait for both the website to fully load and an extra period of time before visiting another website. Consequently, we emulate the behavior of a normal user that stumbled upon a website. Therefore, we make concerted effort not to perform any type of DoS attack to the visited website.

In accordance to the GDPR and ePrivacy regulations, we did not engage in collection of data from real users. Also, we do not share with any other entity any data collected by our crawler. Moreover, we ensure that the privacy of publishers and website owners is not invaded. To that extent, we do not collect any of their information (e.g., email addresses) and only discuss names of websites and not the legal entities that control them or their administrators, as we did in Chapter 5. Last but not least, we intentionally do not make our dataset public, to ensure that there is no infringement of copyrighted material.

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 *ACM CCS*, 2014.
- [2] Pushkal Agarwal, Sagar Joglekar, Panagiotis Papadopoulos, Nishanth Sastry, and Nicolas Kourtellis. Stop tracking me bro! differential tracking of user demographics on hyper-partisan websites. In *WWW*, 2020.
- [3] Ram Aliya and Murgia Madhumita. Data brokers: regulators try to rein in the 'privacy deathstars'. <https://www.ft.com/content/f1590694-fe68-11e8-aebf-99e208d3e521>, 2019.
- [4] Waqar Aqeel, Balakrishnan Chandrasekaran, Anja Feldmann, and Bruce M Maggs. On landing and internal web pages: The strange case of jekyll and hyde in web performance measurement. In *Proceedings of the ACM Internet Measurement Conference*, pages 680–695, 2020.
- [5] The Chromium Authors. Chrome devtools protocol. <https://chromedevtools.github.io/devtools-protocol/>, 2014.
- [6] Brave Browser. What's brave done for my privacy lately? episode #3: Fingerprint randomization. <https://brave.com/privacy-updates-3/>, 2020.
- [7] Pamela Bump. The death of the third-party cookie: What marketers need to know about google's 2022 phase-out. <https://blog.hubspot.com/marketing/third-party-cookie-phase-out>, 2021.
- [8] Aaron Cahn, Scott Alfeld, Paul Barford, and Shanmugavelayutham Muthukrishnan. An empirical study of web cookies. In *Proceedings of the 25th international conference on world wide web*, pages 891–901, 2016.
- [9] European Commission. Proposal for a regulation on privacy and electronic communications. <https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-privacy-and-electronic-communications>, 2017.
- [10] Cookiebot. Google ending third-party cookies in chrome. <https://www.cookiebot.com/en/google-third-party-cookies/>, 2021.

- [11] CookiePro. The cookie law explained. <https://www.cookieelaw.org/the-cookie-law/>, 2011.
- [12] Joseph Cox. Leaked documents expose the secretive market for your web browsing data. <https://www.vice.com/en/article/qjdkq7/avast-antivirus-sells-user-browsing-data-investigation>, 2020.
- [13] A. Dabrowski, G. Merzdovnik, J. Ullrich, G. Sendera, and E. Weippl. Measuring cookies and web privacy in a post-gdpr world. In *PAM*, 2019.
- [14] 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. *arXiv preprint arXiv:1808.05096*, 2018.
- [15] DPO LLC. Article 4 - recital 30. <http://gdpr-text.com/read/article-4/>, 2020.
- [16] Sam Dutton. What is floc? <https://web.dev/floc/>, 2021.
- [17] Peter Eckersley. How unique is your web browser? In *PETS*, 2010.
- [18] Rob van Eijk, Hadi Asghari, Philipp Winter, and Arvind Narayanan. The impact of user location on cookie notices (inside and outside of the european union). In *Workshop on Technology and Consumer Protection (ConPro'19)*, 2019.
- [19] Hazem Elmeleegy, Yinan Li, Yan Qi, Peter Wilmot, Mingxi Wu, Santanu Kolay, Ali Dasdan, and Songting Chen. Overview of turn data management platform for digital advertising. *Proceedings of the VLDB Endowment*, 6(11):1138–1149, 2013.
- [20] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *ACM CCS*, 2016.
- [21] 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*, pages 289–299, 2015.
- [22] Interactive Advertising Bureau Europe. Cmp list. <https://iabeurope.eu/cmp-list/>.
- [23] Interactive Advertising Bureau Europe. Transparency & consent framework. <https://iabeurope.eu/transparency-consent-framework/>, 2018.
- [24] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. Tracking personal identifiers across the web. In *PAM*, 2016.

- [25] David Fifield and Mia Gil Epner. Fingerprintability of webrtc. *arXiv preprint arXiv:1605.08805*, 2016.
- [26] Centre for Advanced Visualisation and Interaction (CAVI) Aarhus University. Consent-o-matic. <https://github.com/cavi-au/Consent-O-Matic>, 2019.
- [27] Imane Fouad, Cristiana Santos, Feras Al Kassar, Nataliia Bielova, and Stefano Calzavara. On compliance of cookie purposes with the purpose specification principle. In *IWPE 2020*, 2020.
- [28] Alex Hern and Jim Waterson. Sites block users, shut down activities and flood inboxes as gdpr rules loom. <https://www.theguardian.com/technology/2018/may/24/sites-block-eu-users-before-gdpr-takes-effect>, 2018.
- [29] Maximilian Hils, Daniel W Woods, and Rainer Böhme. Measuring the emergence of consent management on the web. In *Proceedings of the ACM Internet Measurement Conference*, pages 317–332, 2020.
- [30] Apple Inc. Apple introduces macos mojave. <https://www.apple.com/newsroom/2018/06/apple-introduces-macos-mojave/>, 2018.
- [31] Disconnect Inc. Disconnect - tracker protection lists. <https://github.com/disconnectme/disconnect-tracking-protection>.
- [32] FingerprintJS Inc. Fingerprintjs. <https://github.com/fingerprintjs/fingerprintjs>.
- [33] FingerprintJS Inc. Fingerprintjs pro. <https://web.archive.org/web/20211215185425/https://fingerprintjs.com/>, 2021.
- [34] Google Inc. Puppeteer: Headless chrome node.js api. <https://pptr.dev/>, 2017.
- [35] Costas Iordanou, Nicolas Kourtellis, Juan Miguel Carrascosa, Claudio Soriente, Ruben Cuevas, and Nikolaos Laoutaris. Beyond content analysis: Detecting targeted ads via distributed counting. In *ACM CoNEXT*, 2019.
- [36] John E Dunn. Google chrome to start blocking downloads served via http. <https://nakedsecurity.sophos.com/2020/02/10/google-chrome-to-start-blocking-downloads-served-via-http/>, 2020.
- [37] Jonathan Mervis. Is cookieless tracking the future of web analytics? www.amazeemetrics.com/en/blog/is-cookieless-tracking-the-future-of-web-analytics/, 2020.
- [38] Erin Kenneally and David Dittrich. The menlo report: Ethical principles guiding information and communication technology research. *Available at SSRN 2445102*, 2012.

- [39] Balachander Krishnamurthy and Craig Wills. Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the 18th international conference on World wide web*, pages 541–550, 2009.
- [40] Pierre Laperdrix, Benoit Baudry, and Vikas Mishra. Fprandom: Randomizing core browser objects to break advanced device fingerprinting techniques. In *International Symposium on Engineering Secure Software and Systems*, 2017.
- [41] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints. In *IEEE S&P*, 2016.
- [42] Hoan Le, Federico Fallace, and Pere Barlet-Ros. Towards accurate detection of obfuscated web tracking. In *IEEE M&N*, 2017.
- [43] P. Leach, M. Mealling, and R. Salz. A universally unique identifier (uuid) urn namespace. <https://datatracker.ietf.org/doc/html/rfc4122>, 2005.
- [44] California State Legislature. California consumer privacy act of 2018 [1798.100 - 1798.199.100]. https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5, 2018.
- [45] C. Matte, N. Bielova, and C. Santos. Do cookie banners respect my choice? : Measuring legal compliance of banners from iab europe’s transparency and consent framework. In *IEEE S&P*, 2020.
- [46] Jonathan R Mayer and John C Mitchell. Third-party web tracking: Policy and technology. In *IEEE S&P*, 2012.
- [47] Maryam Mehrnezhad. A cross-platform evaluation of privacy notices and tracking practices. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 97–106. IEEE, 2020.
- [48] Keaton Mowery and Hovav Shacham. Pixel perfect: Fingerprinting canvas in html5. *Proceedings of W2SP*, 2012.
- [49] Mozilla. Version 69.0, first offered to release channel users on september 3, 2019. <https://www.mozilla.org/en-US/firefox/69.0/releasenotes/>, 2019.
- [50] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. Privaricator: Deceiving fingerprinters with little white lies. In *WWW*, 2015.
- [51] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *IEEE S&P*, 2013.

- [52] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.
- [53] Lukasz Olejnik, Gunes Acar, Claude Castelluccia, and Claudia Diaz. The leaking battery. In *Data Privacy Management, and Security Assurance*, pages 254–263. Springer, 2015.
- [54] Michalis Pachilakis, Panagiotis Papadopoulos, Nikolaos Laoutaris, Evangelos P Markatos, and Nicolas Kourtellis. Measuring ad value without bankrupting user privacy. *arXiv preprint arXiv:1907.10331*, 2019.
- [55] Michalis Pachilakis, Panagiotis Papadopoulos, Evangelos P Markatos, and Nicolas Kourtellis. No more chasing waterfalls: a measurement study of the header bidding ad-ecosystem. In *Proceedings of the Internet Measurement Conference*, pages 280–293, 2019.
- [56] Emmanouil Papadogiannakis. Consent guard. <https://gitlab.com/papamano/consent-guard>, 2021.
- [57] Elias P Papadopoulos, Michalis Diamantaris, Panagiotis Papadopoulos, Thanasis Petsas, Sotiris Ioannidis, and Evangelos P Markatos. The long-standing privacy debate: Mobile websites vs mobile apps. In *ACM WWW*, 2017.
- [58] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *WWW*, 2019.
- [59] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. The cost of digital advertisement: Comparing user and advertiser views. In *WWW*, 2018.
- [60] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay to reach you? In *IMC*, 2017.
- [61] The European Parliament and the Council of the European Union. Directive 2002/58/ec of the european parliament and of the council of 12 july 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (directive on privacy and electronic communications), 2002.
- [62] 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), 2016.
- [63] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *NDSS*, 2019.
- [64] Jordan S Queiroz and Eduardo L Feitosa. A web browser fingerprinting method based on the web audio api. *The Computer Journal*, 62(8):1106–1120, 2019.
- [65] Philip Raschke and Axel Küpper. Uncovering canvas fingerprinting in real-time and analyzing ist usage for web-tracking. In *Workshops der INFORMATIK 2018-Architekturen, Prozesse, Sicherheit und Nachhaltigkeit*, 2018.
- [66] Caitlin M. Rivers and Bryan L. Lewis. Ethical research standards in a world of big. *F1000Research*, 3, 2014.
- [67] 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 *ACM Asia CCS*, 2019.
- [68] Rebecca Sentance. Gdpr: Which websites are blocking visitors from the eu? <https://econsultancy.com/gdpr-which-websites-are-blocking-visitors-from-the-eu-2/>, 2018.
- [69] Than Htut Soe, Oda Elise Nordberg, Frode Guribye, and Marija Slavkovik. Circumvention by design-dark patterns in cookie consent for online news outlets. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–12, 2020.
- [70] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. TALON: An automated framework for cross-device tracking detection. In *USENIX RAID*, 2019.
- [71] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. Clash of the trackers: Measuring the evolution of the online tracking ecosystem, 2020.
- [72] Nick Statt. Apple updates safari’s anti-tracking tech with full third-party cookie blocking. <https://www.theverge.com/2020/3/24/21192830/apple-safari-intelligent-tracking-privacy-full-third-party-cookie-blocking>, 2020.
- [73] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Beyond the front page: Measuring third party dynamics in the field. In *Proceedings of The Web Conference 2020*, pages 1275–1286, 2020.

- [74] 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 preprint arXiv:1811.08660*, 2018.
- [75] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (un) informed consent: Studying gdpr consent notices in the field. In *CCS*, 2019.
- [76] Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy. Fp-stalker: Tracking browser fingerprint evolutions. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 728–741. IEEE, 2018.
- [77] John Wilander. Intelligent tracking prevention 2.3. <https://webkit.org/blog/9521/intelligent-tracking-prevention-2-3/>, 2019.
- [78] Marissa Wood. Today’s firefox blocks third-party tracking cookies and cryptomining by default. <https://blog.mozilla.org/en/products/firefox/todays-firefox-blocks-third-party-tracking-cookies-and-cryptomining-by-default/>, 2019.