Characterising and Protecting Against Web Trackers Across The World

Hu, Xuehui

Awarding institution:
King's College London

Characterising and Protecting Against Web Trackers Across The World

Xuehui Hu

Supervisor: Dr. Guillermo Suarez de Tangil
Prof. Nishanth Sastry

Department of Engineering
King's College London

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2022

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been previously included in a thesis or dissertation submitted to this, or any other University. This dissertation represents my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Xuehui Hu
September 2022

Acknowledgements

As it comes to the completion of this dissertation, I am overwhelmed with emotions. Looking back, I need to express sincere gratitude and the guidance I received during the four-year research.

First, I would like to thank my supervisors *Prof. Nishanth Sastry* and *Dr. Guillermo Suarez de Tangil* for the kindly mentor. These four years have been very meaningful for me, not only have I learnt professional knowledge here, but I have also been able to improve my abilities in various aspects. And the help and guidance of *Prof. Mainack Mondal*, which greatly deepened my understanding of privacy hazards caused by human factors. Their profound knowledge, rigorous work attitude to research, and approachable personality have always resharpened and infected me, and will be a valuable asset for the rest of my life. Not only helped me to sharpen my ideas and meticulously advise me academically, but they also created a good academic atmosphere.

Except that, it is said that life is a long road, I cherish each opportunity I obtained in King's College London and each team member I met in *Netsys*. Thanks to Aravindh Raman and Sagar Joglekar for their initial guidance, Pushkal Agarwal for sharing both the study and life, Abdullahi Abubakar for some help with back-end techniques, and each gathering with all the members. Talking to our *Netsys* group allows me to think outside the box when it comes to finding clarity. The encounter with the team is about to become the best memories of the PhD experience.

Subsequently, I want to express my gratitude to my parents who support my pursuit of research work as well as my daily enlightenment. Your inspiration, infinite tolerance and understanding have been my greatest motivation, always by my side, and I would not be where I am today without you. I wish you happiness and good health.

Last but not the least, I would like to express my heartfelt gratitude to all my family, tutors and friends who have cared for and encouraged me.

I hope this paper will not be the end of academics, but the beginning of my maturity.

Abstract

Our digital world is awash with cookies, which are simple text files that keep website specific states on the web browser, such as auto-filled login fields. Although cookies are inherently harmless, third-party vendors use the tracking capability for commercial profits, e.g., cookie matching for audience targeting. This dissertation analyses browsing behaviour based on a large real user dataset collected by a browser extension developed, containing data of 2,537 users from 106 countries until August 2021 (from 10k+ installers). Then, providing solutions to inhibit third-party sharing/profiling and automated cookie protection tools.

The first part studies the third-party ecosystem in different countries, revealing the impact of the type of first-party website sectors and the location of the user on the number of third parties in the wild. Results demonstrate that most users who are interested in a given site category are likely to encounter category-specific third parties, and around 65% of re-visited websites tend to offer more third parties to the same user profile. In terms of the user location, China is prone to a home-grown third-party ecosystem compared with the UK, due to China Great Firewall's access blockade of top third parties (i.e., Google, Facebook,etc.).

To better understand the usage of cookies, I utilise the Cookiepedia database as the ground truth for a four-way classification (i.e., strictly necessary, performance, functionality and targeting/advertising cookies). The machine learning-driven framework achieves 94% F1 score and 1.5 ms latency, only 9.79% and 13.35% in the real-user dataset are identified as necessary and functional cookies. Briefly, most cookies are beneficial to the website rather than the user experience.

After the preliminary analysis on the status quo, the dissertation proposes solutions to restrict cookie-based tracking for online behaviours from two aspects. One is a management assistant for multi-account containers for the reduction in third-party interconnectivity based on common third parties in browsing histories (i.e., "tangle factor"). Evidence shows that removing top third-party vendors does better than all ad blockers in decreasing interconnectedness. And uBlock origin is the best one among ad blockers, reducing the raw number of third parties by 60% and required containers by 40%.

The other solution is the auto-processing of the GDPR minimal data option. Since May 26, 2018, the General Data Protection Regulation (GDPR) was promulgated in the EU to protect personal data without user approval. By the end of 2018, third-party cookies of UK users drop by over 10%. However, the consent fatigue and lack of an automatic consent setting mechanism resulted in the rebound of third-party cookies in 2019. Therefore, I build and deploy a browser extension to automatically assist users to protect user privacy in 85% of the websites with GDPR notices, reducing targeting/advertising cookies by 44.6%.

Concisely, this dissertation mainly addresses the collection and classification of real-time browsing data in the wild, privacy risks of the third party interconnected tunnels and the lack of an automated GDPR-enforcing mechanism. And the field deployments increase the feasibility and usability, successfully hardening the protection against user privacy while browsing and paving the way for the automated global online privacy protection.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The Internet is the first thing that humanity has built that humanity doesn't understand, the largest experiment in anarchy that we have ever had

—— Eric Schmidt, 1955

1.1 Introduction

The Internet initially provides individuals with access rights that were not controlled by rules or laws but with barriers to transparency. In recent years, the rapid development of the Internet has promoted an increase in the generation and sharing of online personal data. The existed regulatory and policy protection are constantly being challenged. In the post-digital era where digitisation becomes commonplace, as all organisations develop digital competency, and technological innovation or improvement is no longer the pursuit of majority [Cra13]. Due to the evolving digital competitive fields, it is more necessary to obtain the differentiation advantages than ever. Technology-driven interactions [DCB19] becomes the next milestone, establishing the ongoing, targeted relationship with users to maintain the nonstop connection.

Among that, the convenience proposed by web cookies technology is irresistible, unobtrusive and ubiquitous, involving auto-filled login fields, hassle-free shopping services, language translation, etc.

Web cookies, described in basic terms, are small blocks of data created by a web server and used by a device while the user is browsing sites. Cookie files could be used to remember the user's state attributes during each browsing session, which is then available to the website and the relevant tracking company to understand the user's identity or to determine the user's behaviour. Since the inception of cookies in 1994 [GM94], cookies (especially third-party cookies), have become ubiquitous, resulting in a growing trend towards misuse of user privacy in recent years. Generally, cookies are used at the application layer, but they constantly leak the massive amount of information that is used for commercial purposes according to users' browsing activities.

For example, the precise ad/targeting technology that underpins the online ad business often relies on third-party cookies. Third-party vendors could conduct personalised in-depth analysis according to the collected cookies, precisely targeting audiences according to the demand of advertisers. Due to the disempowerment on the awareness of online privacy, increasing third-party vendors takes advantage of it for dataveillance, shareveillance and deliberately opaque personal data collection [Bir17]. **While improving digital efficiency, it comes at the expense of the user's privacy.** But cookies for online users mainly involve two different types, which have distinct flavours and privacy risks: first-party cookies and third-party cookies [CABM16b]. Because the third-party cookies are placed by external partners and commonly used for profit, third-party cookies are prone to being surveilled and targeted [Kes05].

Governments prevent the unauthorised transmission of personal data through a sequence of privacy regulations, such as GDPR in EU [Cou16], CCPA in US (California) [Sta18], PIPEDA in Canada [Par00], LGPD in Brazil [Pre18], etc. Since 2018, these policies have already exerted an imperceptible influence on website vendors and individual users to a certain extent, deepened the understanding of the necessity of online privacy, awareness of data protection, and privacy policy enforcement. However, the undue reliance on the user approval ("consent") for the legal data processing greatly promotes the generation of "dark patterns" under regimes.

According to the above two scenarios, there are observations that:

1. The number of trackers per site is affected by the popularity and category of the website, as well as the country-specific third parties.
2. Due to the lack of attention to niche interest site, Alexa based studies are systematically over estimating the amount of tracking that individual users may experience.
3. The protection provided by ad blockers tends to decrease the number of independent third parties rather than the interconnectivity of the third-party ecosystem (i.e., the chromatic number of website cliques that built with first-party website nodes and shared third-party edges between nodes).
4. For websites that provide GDPR notices, accepting the default choices typically ends up storing more cookies on average than before GDPR.
5. The lack of adequate cookie classification coverage and automated privacy policy enforcement mechanism resulted in GDPR's failure to sufficiently convert enhanced user privacy awareness into better privacy protection.

1.2 Thesis statement

These observations describe the pervasive profiling of web trackers and the demand for tightening online privacy. That provokes the following research subjects of this dissertation (listed in order of project sequence rather than importance):

How to feasibly characterise web trackers into the user-understandable ecosystem and *programmatically* enforce GDPR to prevent third-party tracking that are unrewarding to users?

In the practice of problem-solving, I first formulate an ecosystem model for the context of first-/third-party network scale data. For concreteness, I study the first-/third-party match-up correspondence by examining the browsing history data collected from Thunderbeam users (§3.1.1). Observations based on the four-layer model introduce the first problem:

Problem 1 *How does the website's dynamic strategy affect the number of trackers through the popularity, category, and location in-the-wild? Is there traffic discrimination?*

The number of third-party trackers on each website are generally used to ascertain leakage risks of one given first-party website due to the penetration of third-party trackers.

Considering that differences in regulatory regimes, the geographical location of users might pose an influence on the ability to track. In 2004, the survey of [BJKL04] has examined other possible explanations of differences in network privacy concerns: cultural values, Internet experience and political organizations. Then, the study [FHUM14a] from Falahrastegar et al. inspects the existence of large and small third-party services in each region, both globally and regionally, and whether these services penetrated into other areas. In 2016, Suphannee et al. [SPK16] audited the range of tracking services and evaluated the degree that online users suffered from the cookie hijacking attack and the number of exposed accounts in various services. Top3 in ranks are Doubleclick, Google and Amazon and two services of them are banned in China, which could be the reference to the possible cause of the lower number of China-based trackers. Except that, authors in a 2019 study [SM19b] note that the prominence of third-party trackers changes

significantly with the geographic location, providing websites from as many as 56 countries in different political, cultural, etc.

However, Alexa domains, crawling by OpenWPM, might encounter an inherent limitation of server IPs, as well as the lack of diversity. For a given set of visited sites, this dissertation measures the violation of online privacy based on the number of containers required to isolate different first-party websites after interrupting the shared third-party "edges". **The value of real users' third party data is the focus in this research.**

Although there are extensions like multi-account containers offer ways to separate third (and first) parties from each other by creating different cookie stores for different contextual identities, the manual allocation of various containers is necessary. Therefore, the next part focuses on the characteristics and quantification of the third-party interconnectedness of a set of first-party websites:

Problem 2 *If the above policies are uniformly applied to different sets of websites, how many containers will be required to separate different first-party websites that share one or more third parties?*

Another important notion to the arsenal of technologies for preventing tracking of users is the *multi-account containers,* often referred to simply as "containers". Pioneered by Firefox [Moz15], containers are a way to separate different sets of cookies from each other. Containers were initially intended for providing "contextual identities"[1], i.e., to create different user identities depending on the context of operation. For example, containers allow a user to cleanly separate their work and personal identities on the same browser. Different containers can also be used to login to the same sites with multiple user ids (for example, users having several email accounts from the same provider can be simultaneously logged in to each of them in different containers). In this sense, containers are similar to other browser extensions such as

[1] APIs for contextual identities: https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/Work_with_contextual_identities

Multifox [Mar06] or CookieSwap [Ext19] (or the similar Swap my Cookies extension on Chrome [FDe13]).

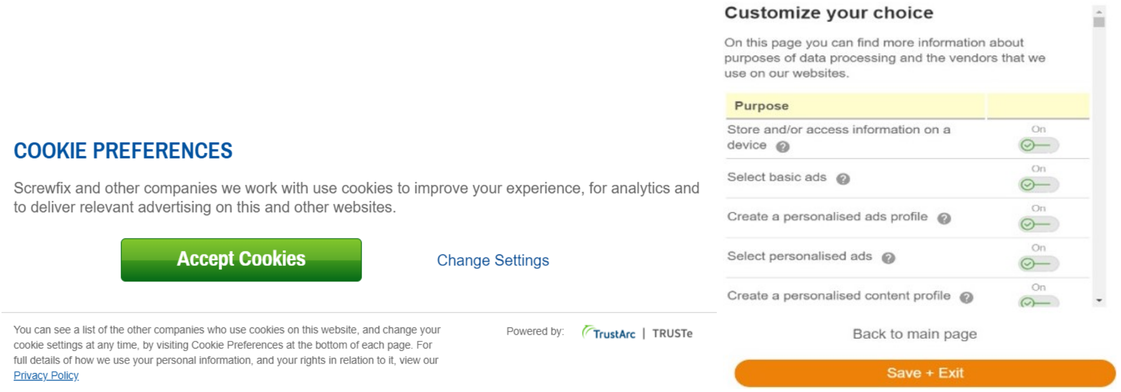
Firefox containers essentially create a different user profile within each container, providing a different database of cookies and storage for each[2]. Thus, each container identity is kept separate from the others, and information such as third party cookies are not shared across containers. Firefox suggests [Moz15] that this can be used to also achieve additional privacy, for instance by placing a user's shopping websites into a separate container from financial websites such as banks and credit cards.

In this research, "containers" are used to decrease interconnectivity and improve privacy of the third-party ecosystem. Compared with independent first-/third-party communities, websites that share connections through at least one common third parties more focus on the contextual identity of the third-party tracking model. Therefore, the number of required containers is according to the number of first-party pairs that needs to separate in the browsing history by user or country.The larger the number of containers required, the lower the interconnectivity of the browsing history. Unlike traditional measure violation of privacy, websites that can co-exist with no privacy loss in the container means that untangling interconnected sites is another method of privacy protection.

In addition to efforts on minimising cookies **through the isolation of third parties with "containers"**, the governments also promulgate Internet privacy acts by country to protect users from third-party tracking and data leakage. This dissertation focuses on the EU's General Data Protection Regulation (GDPR), which demands no data collection before user approval. The investigation of GDPR cookie notices on first-party websites motivates an extended cookie classification database and automated data minimisation for GDPR that I developed in this dissertation.

[2] In practice, a `userContextID` column is added to the cookie database, and only cookies matching the context ID of the container are sent to the website.

Problem 3 *How can I properly identify the purpose of a cookie that is not catalogued in the current cookie database during a user's web browsing? And help users automatically manage GDPR consent settings?*



COOKIE PREFERENCES

Screwfix and other companies we work with use cookies to improve your experience, for analytics and to deliver relevant advertising on this and other websites.

You can see a list of the other companies who use cookies on this website, and change your cookie settings at any time, by visiting Cookie Preferences at the bottom of each page. For full details of how we use your personal information, and your rights in relation to it, view our Privacy Policy

(a)

Customize your choice

On this page you can find more information about purposes of data processing and the vendors that we use on our websites.

Purpose	
Store and/or access information on a device	On
Select basic ads	On
Create a personalised ads profile	On
Select personalised ads	On
Create a personalised content profile	On

(b)

Figure 1.1: Cookie consent notices presented to the users in GDPR-compliant websites.

A majority of the websites in the EU show users consent notices, to be compliant with GDPR and the ePrivacy or "cookie" directive. However, these consent notices are often presented to end-users as a barrier to access the website content[CDL+18]. In addition, websites commonly employ dark patterns to nudge users into agreeing to be tracked[MAF+19]: For example, the "Accept cookies" or other default setting option may be selected by default (Figure 8.1(a)), and even when privacy-conscious users click the non-highlighted "Cookies Settings" option, they may be inundated with too many choices (Figure 8.1(b)), causing decision fatigue. Fixing such **"user interface tricks"** is an active topic of research. Even if these user interface design problems are addressed, cookie consents still have several key problems which affect their usability. The three sub-problems related to cookie consent notices are as follows:

Consent fatigue. Under GDPR, data controllers are required to obtain consent before using a data subject's (i.e., user's) personal data. In practical terms, this means that *each* website that a user visits throws up a separate consent notice which the user has to navigate carefully in order to protect their privacy. The huge diversity in the designs of GDPR consent notices places additional cognitive burden on users. This sets up the conditions for so called "consent fatigue" and many users end up clicking on the "I accept" button [HH19]. Other research has shown that such privacy fatigue has "a stronger impact on privacy behaviour than privacy concerns do, although the latter is widely regarded as the dominant factor in explaining online privacy behaviour" [CPJ18].

Lack of consent withdrawal mechanisms. To avoid annoying the user on every visit, websites typically ask for consent to collect users' data only on the initial visit, and the consent notice banners presented on this initial visit set one or more cookies recording the user's initial preferences regarding data collection and also record the fact that the notice has been served. A side effect of this is that on most websites today, once the users agree to be tracked, it is exceptionally hard to change their settings, or even to review the consent notice after giving consent once. This is in direct violation of Article 7 of GDPR which gives data subjects the "right to withdraw consent". Today, the only robust (and prohibitively difficult) solution to change the consent settings or to withdraw consent entirely is to explicitly identify the handful of cookies stored in users' browser for a particular website that record the user consent, and to remove/modify them manually. This is well beyond the reach of most ordinary netizens.

Cookie consents globally. A recent work designed, built and deployed a Chrome plugin[HdTS20] in-the-wild to show connections between first party websites and third party tracking services used by them. According to the Chrome webstore[3], this plugin appears to be actively used by more than 10k users. Their data shows that tracking is rampant in most countries including outside the EU and many countries such as China have home grown third-party ecosystems that are not well covered by popular ad blockers[HS20, HdTS20]. I attribute this strong violation of user privacy at global scale partially to the absence of cookie consent notices — with a few exceptions such as GDPR & CCPA in California, most jurisdictions do not even have the protection

[3] "Thunderbeam-Lightbeam for Chrome": tiny.cc/lightbeam-chrome-plugin

of the cookie consent framework. These privacy violations provide a strong need to extend the privacy protection of cookie consents globally.

To address the above issues, cookie consent notices work by actually setting a small set of additional cookies whose values record users' preferences on what data they permit the websites to collect, whether for tracking or other purposes. There is a simple and straightforward goal to figure out whether can *programmatically* set the right values of GDPR notices to prevent tracking.

1.3 Thesis Arguments

The problems stated above would be approached by observations and comparisons of the browsing ecosystem models. As the user-centric browsing Internet, I deploy the browser extension for collecting real-user browsing histories, rather than just selectively studying on the scale of top websites. That is not only convenient for field deployment and participant recruitment but also more cost-effective than using website embedding.

The thesis is that third-party ecosystem are used to tailor tracking strategy to the user base, which could be effectively and friendly restricted by browser extensions.

I substantiate the thesis by deploying plugin-based solutions to the two aspects posed above. Privacy-enhancing Web browser extensions were developed independently of each other, **yielding useful insights** of tackling tangles between the user's browsing histories and the web trackers. I first gain observations through the datasets collected from browser extension users to characterise the first-party nodes and third-party tracking edges in the web journey. The structure of this graph-theoretic network is then used to develop **restriction methods against third-party tracking (e.g., Tangle Factors and CookieCutter extension).**

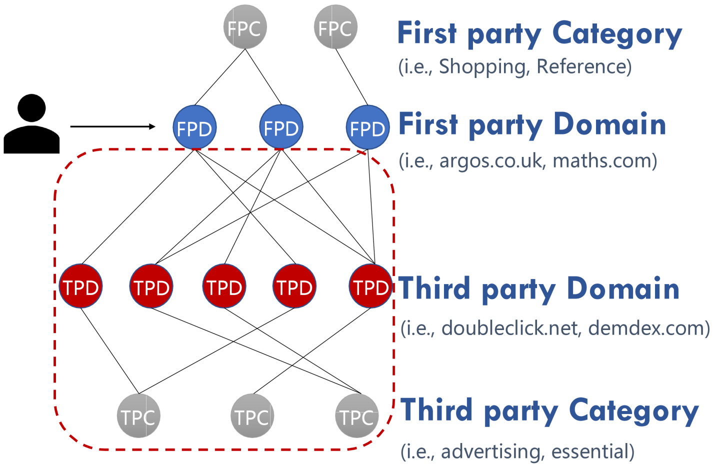

FPC
FPC
First party Category
(i.e., Shopping, Reference)
FPD
FPD
FPD
First party Domain
(i.e., argos.co.uk, maths.com)
TPD
TPD
TPD
TPD
TPD
Third party Domain
(i.e., doubleclick.net, demdex.com)
TPC
TPC
TPC
Third party Category
(i.e., advertising, essential)

Figure 1.2: The user visits First Party Domains (FPD) belonging to different First Party Categories (FPC), then these FPD simultaneously load trackers from Third Party Domains (TPD) of different Third Party Categories. Noted that There are 16 FPC and 8 TPC in the database.

1.3.1 Visualised third-party ecosystem

Considering the fact that Chrome occupies nearly 70% of the usage share held by desktop browsers [Sta21a] and the lack of privacy visualization tools for Chrome, this study focuses on the analysis of the third-party privacy ecology of Chrome's user base as a preliminary exploration.

To interactively visualises users whith the cookie third-party ecosystem and understand how third-party sites learn from users' browsing behaviours, I develop the Chrome extension, called "Thunderbeam"(see §3.1.1). It offers an alternative version of Lightbeam [Moz12] for Chrome users, namely, a graph-based visual representation to explain how first-party websites visited by the user introduce trackers. This extension births from Lightbeam [Moz12] in Firefox [Moz12], now it supports related features based on Chrome's APIs and adds additional functionality. The specific challenges that need to be tackled for tne deployment are described in §3.1.1.

The model in Figure 1.2 describes the graph-based third-party tracking mechanism among the user browsing histories It allows the developer to extract the relationship between the first-party domain (FPD) and third-party domains (TPD) the user visits. However, because the same third party might be loaded by multiple first-party websites, the third party domain is in a unique privileged position. Therefore, from the frequency statistics of the tracking model, it make users more easily understand the first-party browsing habits over time or the difference between cohorts of users. It also better connects the changes of the third party with the first party.

In terms of functionalities, the four main benefits of Thunderbeam are as follows:

- **Connectivity**: Throughout Thunderbeam, users will be able to explore how many third parties are aware of their browsing habits. Thunderbeam will display first-party websites using rounded circles and third-parties using triangles. The system will start connecting first-parties with third-parties as websites drop cookies in real-time. According to the number of nodes and the degree of connection, users will become aware of how intertwined "their" Web ecosystem is (i.e., how connected are the local and global websites they visit).

- **Categorisation** of first/third parties stored in browser: The extension provides two unique databases that I leverage to characterise the the respective first-party and third-party websites that are visited. The first-party database has been collected from Alexa top500 sites listed for each category [4]; and the third-party database is generated by merging eight different lists from a wide range of sources. The collection methodology prioritises accuracy and thus I use manually classification to annotate third parties which are not in any of the known lists (see [HdTS20] for more details).

- **Analytics**: There are anonymous donations collected from other users of Thunderbeam. This could showcase to the audience a unique snapshot of how the Web ecosystem works around the globe (see https://tiny.cc/ThunderbeamVis). Meanwhile, the research is ethical by following the guidelines from The Belmont Report [Bea08]. First, it does not request personally identifiable data such as name, nor does it collect information that could be used to identify users like

[4] Alexa by category: www.alexa.com/topsites/category

IP addresses. Instead, it would only collect a consistent UUID stored in the extension to represent the user profile. Also, it does not collect any sensitive information (e.g., age or gender) except for the browsing history. To realise the automatic weekly data collection for the academic research, a toggle switch is created on the left side of the extension dashboard to make users easier to opt-in/out of the data collection consent. And it is to guarantee users: 1) willingly share their anonymised browsing history with the server, 2) have been offered the right of withdrawing from the collection.

- **Offline targeted analysis**: Users of Thunderbeam are able to download their own private data file and run it through an analysis suite purposely of design and implemented for Thunderbeam. As a result, users will be able to analyse their own local data. This will work as follows: 1) participants will be given access the Data Visualisation and Analysis framework (Figure 3.1(a))) they will generate a .json file through the button *YourData* (Figure 3.1(b))) they will be given instructions on how to load the data into the aforementioned analysis framework (Figure 3.1(c)). As part of the analysis suit, a widget would be provided to perform third-party ranking based on the anonymised and aggregated data that have collected from volunteers.

Traditionally, third-party tracking tools are not directly used to assist in making systems-level decisions, but as an approach to either understand the third-party content or accept its implicit coercive tendencies. For instance, Disconnect for displaying online browsing behaviours, Ghostery and Adblock Plus blocking their respective defined trackers. Users seldom contact the specific categorisation and shareveillance between third parties in detail, while the newly-developed extension collects both categorical data and behaviour preference of the browsing history.

Third-party disambiguation. When evaluating the exposure of a single first-party website to the user's browsing history in Thunderbeam, it is insufficient to use a third-party domain name. Because a single entity is possible to use multiple domain names, or explicitly hide the degree of tracking, or as a result of discrepancies arising in domain name usage (partly due to mergers). For instance, there are multiple third-party domain names (`doubleclick.net` and `google-analytics.com`) belonging to Google, in this case it is necessary to avoid the duplicate statistics in third-party measurement. In order to eliminate the ambiguity of this situation, I follow the previous work [KW09] and eliminate duplicate third parties controlled by the same authoritative DNS server (ADNS).

Concerning the long-term browsing analysis on wild users, there are dynamic strategies for third-party allocation that are segmented based on the country where the user connects from and the website category that the user connects to. Furthermore, due to the lack of attention to niche interest sites, prior research on Alexa top-ranking lists may systematically overestimate the amount of tracking that individuals may experience. The data-driven analysis suite against the browsing data files could also measure the realistic tracking status of individuals in real-time.

In some cases, sites are not able to support the same set of third-party lists. For example, the Great Firewall (GFW) of China blocks services such as Facebook, Twitter, and Google. Therefore, it demonstrates that a set of country-specific third parties loads in specific countries when users access the same set of first-party websites. Based on the analysis of data collected from the UK/China-located users, it shows that *whether because of user demographic characteristics inferred based on locations or because of GFW, users in China are subject to lesser tracking than users in the UK, even visiting the same set of websites.*

In terms of the third-party records of the same user who stayed in the same country, according to their browsing preferences or interests analysed by the website owner, the third-party platform would also provide a different set of third parties for the same user each time they visit. But as the number of visits increases, the proportion of new third parties would decline. Also, based on the comparison between UK and China users, it is observed that UK users would have a more fast-developing third-party ecosystem over time than China users. There are more insights into the methods described above in §4.

1.3.2 “Tangle Factor” of Interconnectivity Graph

To quantify how third parties and first parties interconnect with each other, this dissertation introduces a new metric called “Tangle Factor”. The metric serves the important purpose of quantifying tangle factors of different users or countries, separating different first-party websites in users’ browsing histories by user or country. Therefore, the larger the number of containers required, the lower the interconnectivity of the browsing history.

I deploy four popular ad blockers (i.e., uBlock origin, Adblock Plus, Ghostery and Adguard) to examine the performance in getting the tangled sites untangled in turn to seek the most effective one in decreasing interconnectivity and improving privacy. However, results depict that while uBlock origin has the best reduction performance among four ad blockers in terms of both Tangle factors (45%) and third party quantity (60%), the removal of popular third parties based on browsing history would be a more effective method of restricting third-party interconnections..

Generalising from the description, any third party can be prevented from learning the (partial) browsing histories of users if two first parties sharing the same third party are placed in different containers. To understand how the isolation of the multi-account containers can prevent tracking, consider the following three-site example, which is pictorially depicted in Figure 1.3. The websites in green and red share one or more third parties (e.g., Facebook, Google DoubleClick etc.), and therefore need to be placed in separate containers; otherwise the common third parties (e.g., Google’s DoubleClick) will be able to infer that the same user visited both the green and red websites. However, the blue site does not share any third parties with either the green or red website and therefore can be placed in the same container with either of those two websites.

Such a graph can be used to answer Problem 2 in an interesting way. The number of required containers provides us with a way to characterise and quantify the interconnectedness of the third-party ecosystem for a given set of first-party websites.

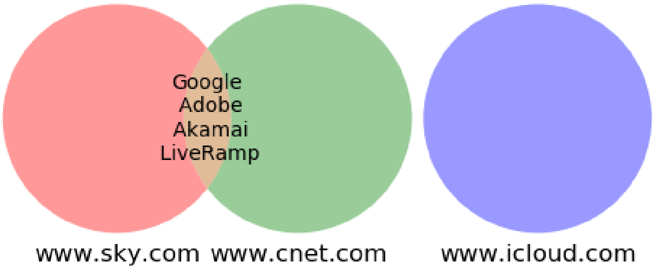


Figure 1.3: Overlapping first parties which share a third party tracker must be placed in separate containers, thus the red and green sites must be separated. The blue website does not share any third parties with either red or green and can be placed together with either of them in a container.

It is termed as the *Third Party Tangle Factor*, or simply the **Tangle Factor** of that set of websites. The higher the **Tangle Factor**, the more the interconnectedness of third party cookie ecosystem.

The tangle factor is calculated by modelling the set of FPs as nodes of a graph, drawing edges between two FPs when they share one or more TPs. It is named as the *first party interconnection graph* (FPIG). Two FPs in the FPIG that share an edge (i.e., share one or more TPs) must be therefore placed in separate containers in order to prevent tracking by the shared TPs. If one specific colour is assigned to each container, and label FPs in the FPIG with the colour of the container they are placed in, it is easy to see that the *vertex chromatic number* of the FPIG, i.e., the number of colours needed for nodes or vertices of the FPIG such that neighbouring vertices which share an edge are coloured differently, gives the minimum number of containers needed to effectively separate that set of FPs. This dissertation regards this as the *tangle factor* of a given set of first-party websites. To understand the third party ecosystem from different vantage points, I apply the Tangle Factor metric to three different sets of first-party websites to demonstrats the correlation between the interconnectedness of websites and the actual numbers of third parties.

Next, according to a “what if” scenario where the most common third parties simply did not exist or were prevented from operating (e.g., through ad blockers), it is accessible to assess and compare to obtain the most effective method to restrict the number of containers needed. This forms the basis of the automatic interference scheme for container allocation; and also is applicable to various countries and cultural contexts.

1.3.3 ML-based assignment of cookies

To confirm whether there is a programmatic alternative to cookie consent notices (Problem 3), this project exhaustively examined all the Alexa top 100 websites in the UK. Only 55 of these websites present users with cookie consent banners and ask users for consent to collect their Personally Identifiable Information (PII) data. Cookies that used to remember the option of consents by **Consent Management Platforms** (CMPs) are called as *GDPR Consent Cookies*.

This research compares the number of (non-GDPR Consent) Cookies set by these websites (for tracking, analytics, etc.) when a user manually chooses the most private option by clicking on the cookie consent banner vs. the number of non-GDPR Consent Cookies set when the GDPR Consent Cookies of these websites are pre-populated for the user (***Consent fatigue*** in Problem 3).

Article 3 of GDPR[5] stipulates that the regulation applies to "applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not." Thus, it is expected that websites operated from or controlled by EU-based establishments should obtain consent from all users globally. That is to say, outside of EEA jurisdictions, websites do not require to provide user with cookie consent notices even if they have loaded consent management scripts and libraries, such as collecting specific names and locations of users. However, I notice that manually setting cookies for GDPR consent still triggers GDPR-level protection against tracking. Therefore, I aim to extend cookie consent functionality from a vantage point within the GDPR jurisdiction region to the global areas. the purpose of this research is to enable GDPR consent cookies available in users'

[5] https://gdpr-info.eu/art-3-gdpr/

browsers outside European Economic Area (EEA), in addition to the cases specified in Article 3 of GDPR

To quantify this systematically, I pick four countries as the primentary examination—UK, USA, India and South Africa. The controlled expriment confirms that the number of cookies in most websites accessed from the three countries outside the jurisdiction of GDPR (*i.e.*, USA, India and South Africa) could be effectively decreased. It suggests that modelling a ML-based classifier for the cookie assignment could contribute to global users the similar privacy protection as provided by GDPR in the EU (solving *Cookie consents globally* in Problem 3).

Finally, given that all the above evidence suggests that GDPR Consent Cookies appear to have complete control over website tracking behaviours, I check whether removing the GDPR consent cookies which were previously set offers a way for users to revisit or change their previously set cookie consents. When the user manually removes the identified GDPR consent cookies and refresh the web page, it displays that the GDPR consent banner pops up again, allowing users to set a different value for their consents (solving ***Lack of consent withdrawal mechanisms*** in Problem 3).

1.4 List of Publications

This section mainly lists all the publications during my four-year Ph.D. research period, working with diverse teams of researchers. The author-lead publications contribute to different chapters of this thesis.

1. **Hu, X.** and Sastry, N. Characterising third party cookie usage in the eu after gdpr. *In Proceedings of the 10th ACM Conference on Web Science*, 2019, pp. 137-141.

2. **Hu, X.** and Sastry, N. What a Tangled Web We Weave: Understanding the Interconnectedness of the Third Party Cookie Ecosystem. *In 12th ACM Conference on Web Science*, 2020, pp. 76-85.

3. **Hu, X.**, de Tangil, G. S., Sastry, N. Multi-country Study of Third Party Trackers from Real Browser Histories. *In 2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2020, pp. 70-86.

4. **Hu, X.**, Sastry, N, Mondal, M. CCCC: Corralling Cookies into Categories with CookieMonster. *13th ACM Web Science Conference*, 2021, pp. 234–242.

1.5 Conclusions and chapter outline

In Chapter 2, I make use of the state-of-art literature in similar fields to contextualise my work in the four aspects of online privacy, namely third-party tracking, contextual identity based container assignment, cookie classification and privacy countermeasures (regulations and protection tools). Then, I explore third-party ecosystems that have developed in different countries in Chapter 4, and evidence the presence of the specialisation of trackers and country-specific patterns in tracking. The third-party tracking model is in favour of dealing with a deluge of browsing data and provides insights towards answers to **Problem 1**. Chapter 5 examine how effectively and uniformly apply the multi-account containers to different sets of websites to answer **Problem 2**. It deepens the understanding of contextual identity in practice, as well as the importance of automation assignments highlighted. A quantitative metric "Tangle Factor" is defined to measure the common third parties shared among a set of visited websites and the effectiveness of countermeasure to interconnections. Under the background of GDPR released in EU (§7), Chapter 6 and 8 introduces the design of a new machine learning-driven framework and its usable browser deployment, which provide varying degrees of support for quantifying and optimising the impact of ad blockers and privacy regulations. As per the mandate posed by **Problem 3**, I discuss the utility of such a model of support automated assistance for GDPR consent cookies, and draw parallels with other state-of-art for solutions with similar intentions.

Briefly, this dissertation demonstrates how to exploit the web browsing ecosystem developed from across the world for automated third-party privacy protection within and outside the jurisdiction of GDPR. And the main contributions are threefold:

1. Re-contextualisation of third-party tracking model to derive intelligent tracking strategies from different sets of categorical sites visited by real users from different countries.

2. Optimise the manual placement of sites into intelligent automation, promoting a wider range of application scenarios for multi-account containers.

3. Deploy the theoretical framework of optimal GDPR consent settings, accompanied by automated individual decision-making and profiling driven by machine learning.

Chapter 2

Literature Review

Summary

Third-party tracking behind the first-party website visits is prevalent in nearly every corner of the web browsing journey. To demystify the fundamentals of third-party tracking, the emphasis of this chapter is to discuss the definition and prior research pertaining to this dissertation, providing a deep dive into technical methods behind invisible surveillance by vendors that users unintentionally interact with. Each relevant work in this dissertation would specifically describe its relative advantages and weakness in each section where they are explained.

2.1 A Brief History of Web Tracking

Awareness of the relevance between privacy leaks and third-party tracking is not recent, [KW09, KNW11] have started over a decade ago that information is risky through transmission between tandem entities via a universally unique identifier (UUID) or browser fingerprinting where is used to uniquely identify the client. A systematical review of the evolution of the third-party tracking ecosystem by 2012 [MM12] promotes public understanding and policy debates on web tracking. Authors claim that a web measurement is an effective approach to understanding trackers. Over the past decade, most research in web tracking has emphasised the use of quantification towards the third-party tracking phenomena.

One of the three key questions posited by Gomer *et al.* [GRMFS13a] in 2013 is about the characteristics of tracking services. They described the tracking network as a small-world network model due to the highly connected small-group entities.

Since 2015, large-scale research based on the Alexa Top Sites web service began to emerge, and it started to rage as the scale of the data set expands. The 10K dataset [LHFE15] makes use of an existed Firefox extension (introduced in [MM12]) for data collection, however, it is novel in identifying third-party trackers that involve sensitive data. After that, the automatic browser instrumentation is the key to large-scale research based on Alexa top 1 million sites [Lib15, SPK16] and even billions of CommonCrawl-based top sites [SK18]. It [SPK16] also stated the online trackers have "long but thin tail" characteristics, which performed similarly to the 'small-world' model. However, it also reveals the limitations of existing mechanisms that difficult to protect users adequately and the overestimation of cookies by numbers in Alexa-based study. For example, the user's explicit logout does not entirely prevent the follow-up access privileges of the cookie holder. At the same time, Metwalley *et al.* [MTM+15b] characterises the online tracking from a passive angle, which checked the coverage of top 100 trackers placed in the dataset collected by Tstat[1]. However, they do not take account of the influence of targeted tracing strategies of third parties towards specific real user profiles.

Such aforementioned works highlight the importance of third-party websites in leaking personal information and motivate this dissertation on the third-party trackers against users in the wild and approval of their use.

To better approximate the observable tracking among real users' browsing histories, [RKW12, TGM15] proposed new measurements (e.g., Firefox add-on) to respectively investigate the impact of trackers on browsing profiles and personal context, but only small-scale real users are involved. Some [LSW+13a, ERE+15, MLXL16] utilise

[1] TCP STatistic and Analysis Tool: http://tstat.polito.it

American Online (AOL) Query Logs dataset [PCT06] to simulate realistic user behaviours. And conducting behavioural research on a large scale is challenging as it demands substantial resources, which are provided in this dissertation by deploying the browser extension in the wild. Despite the smaller-scale Alexa dataset (compared to millions and billions of top-level site analysis), the developed browser extension here has the strength to collect data impacted by thousands of real user profiles and investigate subsequent cookie changes. Furthermore, this dissertation also provides special attention to the relevance between browsing habits of the given user location (i.e., country) who voluntarily participated in this research and the distribution of country-specific third-party trackers.

Nevertheless, the third-party ecosystem is becoming increasingly complicated and nontransparent to online users. Governments, therefore, start to wither introduce or update privacy protection laws around the year 2018, intending to end the lack of enforcement and sanctions in data protection provisions [Alb16]. Different countries have issued various regulations to restrain the digital data leakage of Internet users. For example, General Data Protection Regulation (GDPR) [gdp] targeted at European residents, California Consumer Privacy Act (CCPA) [ccp] against California residents, etc. **One of** central tenets of these data protection tools is to prohibit the collection of sensitive data without user consent.

Although the relevant regulations are not yet consummated [Zar16a], and those viewpoints that the data governance regime reconfigured with reference to the GDPR are incompatible, to a certain extent, further exacerbated the tension between privacy and data demands [Zar17], this notion of data protection has affected multiple countries. Compared with the time GDPR was first released in 2018, I find that more websites and countries have begun to adopt GDPR-related consent notices for data privacy, indicating that GDPR is becoming mature.

And certain organisations subject to GDPR try to use GDPR interfaces designed in dark patterns to steer [MAF+19], mislead [BEK+16] and even deceive [GCL20] (trick [Wal20]) the approval of users, aiming at that user-tailored third-party services are uninterrupted by GDPR. While third parties and policymakers have prepared shortcuts for obtaining and protecting shared data respectively, it is the attitude to which the targeted audience (users) consent to the privacy notice ultimately marks the difference. Therefore, the factors influencing users' willingness to share data are of paramount importance.

Awareness of the importance of the user willingness to share *different types of* data is not recent, having first been discussed in 2013 by Leon *et al.* [LUW+13]. The majority of users at the time were not enthusiastic about paying for additional services that prevent data collection or advertisements, because websites were considered free and privacy is the right without having to pay. An investigation on the cost and worth of privacy [WS19] also casts light on an interesting finding called "superendowment effect", that is, users are only willing to pay little for the protection of data, but the requirements for relinquishing the right to privacy is much more (e.g., accept data sharing).

Then, the gaining focus shifted to how to ensure parties comply with GDPR. On a wide number of occasions, actual behaviour might differ from the stated and the predicted [NHH07]. Although GDPR aims to protect users to alleviate online privacy risks, actual user behaviour might differ from expectations [NHH07]. For instance, due to redundant privacy clauses, users are generally unwilling to spend extra time to browse the right they would offer and waive [RSS+16]. For a variety of reasons, users might unwittingly consent to unobtrusive sharing [BO20, HP19] or illegal data processing by other parties [Chi18b].

In such a case, GDPR needs a series of solutions to ensure that users and websites adhere to legally binding requirements, which were mainly conducted in three aspects in previous studies: *(i)* in-depth analysis of cookie banners [HZJ+19, KS21, UDF+19, NLV+20, HPW+20, SRDK+19a]; *(ii)* algorithmic verification and evaluation of GDPR compliance [BDH18, LKHF20]; *(iii)* automatic data minimisation in GDPR consent settings [MBS20].

2.2 Large-scale Web Tracking in the browser

The evaluation of browser tracking extent mainly involves two aspects: *(i)* country-level third-party comparison of real users based on the domain-based tracking protection lists, and *(ii)* the correlation between the third-party interconnectedness and first-party browsing patterns. Concerning the former location-based analysis, **Due to the difficulty in tracking real-time dynamic third-party ecosystems and the limited user authorization, few scholars have attempted to conduct empirical analysis on real user datasets. With the browser extension development proposed in this project, users could not only save their previous online journeys as local JSON files of a fixed period in addition to online analysis, but also re-analyse them at any time by uploading files to improve the repeatability of user analysis.** Therefore, the user-friendly extension consolidates the novelty of the dissertation's data collection stage compared with using the Alexa datasets, then the evaluation results depend on the tracked third-party ecosystem model.

2.2.1 Domain-based tracking detection

Some studies identify trackers by the domain-based classification [KNW11, CKB12] and look at the prevalence of third parties across these categories, using the McAfee database [Tru]. The major problem of this dataset is that it only provides the category of the domain and does not care about the popularity of the site. However, the observations of datasets in bulk indicates that without considering the site's popularity can lead to over approximations in the amount of tracking. Other works [BCK+14, LUW+13, EN16] consider different categories of websites, but they are all broad without consideration of key fine-grained categories. Differently from these works, I perform extensive manual validation and increase the size of publicly available lists for Western websites by 12.8% and Chinese websites by 23.4%.

Despite the widespread use of domain filter lists, there is a consensus among some researchers that filter lists still might miss trackers because the manual nature of the blacklists generation makes it extremely hard to continuously update the outdated list of third-party domain names due to ever-changing changes. Therefore, studies on third-party detection introduced new features related to distinguishing (non-)trackers based on behaviour-based [WLZW15] and graph-based detection [ISZ+20, KAH+21], which makes it more prone to automatically generate blacklists. However, both ground-truth datasets originate from the existing filter lists centred on ad trackers, which requires improvement contingent on periodical feedback.

Additionally, the privacy extensions with popular tracker-blocking lists (such as EasyList, EasyPrivacy, Disconnect lists) that are commonly used to automatically detect and block third-party trackers have been confirmed in experiments that can only successfully block up to 25% of tracker detection [CGI+20]. The low inefficiency of private-related extensions leads to the omission of a great number of cookie-based tracking, which inspires the improvement of cookie-based categorisation in §6.

2.2.2 Country-level analysis

First, studies in [BJKL04] and [FHUM14a] in 2004 and in 2014 have focused on distinguishing the location of users when discussing third-party trackers. As opposed to findings here, they both conclude that the location of the user has limited influence in the amount and type of tracking. On the other hand, authors in [SK16a] analyse which trackers are used in different countries. They conclude that the Chinese market is not dominated by the same trackers that are popular in other countries. However, their analysis does not provide deep insights and, more importantly, the type of tracking is not contextualised as this dissertation does.

Second, most studies are based either on visiting specific websites such as the `Alexa` most popular websites (e.g., [EN16, PXQ+19, BRAY17, PSL19, MQS17]), or they may at best artificially construct "personae" by initially visiting a number of websites that represent a particular persona or demographic. (e.g., [BW18, SIIK19, CNS20]). In contrast, results based on real-user browsing behaviour are consistent with previous studies, which offers a confirmation that real users in the wild are affected as researchers previously believed. I also highlight the differences where are observed. For instance, the intensive *social-categorised third parties* tracking China users that displayed than UK users in §4.4. Although authors in [ISPL18a] discuss the cross-region tracking flows based on the data collection from 350 real users, their findings are only aimed at two categories of third parties (i.e, advertising and tracking) classified by the list of AdBlock Plus. In contrast to that, I optimise the domain-based categorisation of third parties (reviewed in §2.2.1) that improves the classification accuracy of a total of eight categories. With the broader classification and the one-year tracing of our UK and China participants, results illustrate that a considerable number (over 30K) of trackers are following UK users, which is far riskier than China users.

2.2.3 Third-Party Interconnections

[PRMF16] also introduced a new indicator to describe the neighbourhood degree of one third-party domain, namely, the number of cross-site connected pages for a given tracker domain. Different from likewise evaluating the risk severity of trackers, this dissertation additionally provides an application deployment environment for this metric.

Vyas *et al.* [VMK17] proposed to extend the same-origin policy by adding a so-called origin attributes field in 2017, and separating cookies from different origin attributes. The newly designed mechanism detects a collision between two first parties automatically if third parties are shared, and can be used to create different origin attributes automatically. Thus, the new method can feed into the origin attributes mechanism.

Origin attributes are used for contextual IDs in Mozilla multi-account containers [Moz17]. Mozilla also introduced a special-purpose container targeted at the control of Facebook trackers [Moz18b]. Inspired by these efforts, and generalisation of observations shows that removing the top 10-20 containers can have a hugely beneficial effect.

2.3 Four-part Cookie Categorisation

This section would be divided into two dimensions: *(i)* prior work on understanding usage of cookies with a focus on third-party cookies; *(ii)* prior work on categorising cookies, which tried to bring transparency into the tracking ecosystem.

2.3.1 Categorical cookies

Cookies are an integral part of the Web and were designed to store and *remember* information across sessions about a particular user visiting a particular site[Net02]. However, cookies today are often leveraged for tracking users across services. These tracking cookies, often set by third-parties, store and commercialise information regarding the browsing habits of users, often without user consent.

In fact, privacy violation by third-party cookies about online tracking has become a common problem today. Aside from academic proposals, there are a number of deployed approaches to detect third-party cookie presence and protect online users from privacy intrusion. For instance, while browsing news [AJP+20] or processing online payment [PPAB16] these cookies are generally placed to trace and speculate on users' online activities at scale. Consequently, a flurry of recent studies attempted to identify and detect these third-party cookies in websites. Many of these studies leverage third-party domain names in cookies to detect third party cookies [SCL18, ZCBZ19]. A few studies also leverage the similarity of source HTML codes of a website [JG18] to identify third-party presence and alert users. However, these methods are often computationally expensive and greatly affects the practicality of real-time detection [MPS+13]. Our study contributes to this line of study by designing *CookieMonster* in §6, a novel machine learning-driven method for scalably categorising cookies.

I provide a complementary machine learning-based approach for cookie categorisation and potential blocking which can be used in conjunction with these list-based approaches. In fact, this new method builds on recent work that used a learning approach using web-traffic data [KAH+20]. It captures invisible trackers missed by filter lists using web-traffic from the user's computer and obtains 90.9% accuracy of detection for the Alexa top 10K websites. The newly introduced approach is complementary to that, as both can be used to identify and potentially block trackers. Moreover, the designed system is primarily dependent on cookie *names* for categorisation (removing the need for more computationally expensive capture and analysis of web traffic). Furthermore, not only can trackers be identified, but also necessary and performance cookies are distinguishable, achieving an accuracy of 94%, significantly more than prior work [KAH+20] for third-party tracker detection. By virtue of using cookie names, it is also feasible to evade anti-ad-blockers—tools that are being developed against ad blockers [GP20, ISQ17] which aims to defeat today's ad/tracker blocking systems by manipulating the webpage source code.

2.3.2 Technical aspect of cookie categorisation

With the advent of the General Data Protection Regulation (GDPR) in the EU, cookie categorisation has become more structured. Currently, the most commonly used classification in English language websites is the one proposed by the UK International Chamber of Commerce (UK ICC) and Information Commissioner's Office (ICO). The UK ICC and ICO catalogue cookies into four broad categories [ICC12]: *strictly necessary* cookies, which are essential for the website's function (e.g., logins, shopping carts); *performance* cookies, which collect analytics information to improve a website's performance; *functionality* cookies which remember user choices such as preferred language or location, allowing personalisation of the website to the user; and *targeting/advertising* cookies, typically placed by third-party advertising networks with the permission of the first party website to profile users and serve them ads.

Cookiepdia [One20], a massive dataset of more than 31 million cookies collected from websites and managed by OneTrust (a company for operationalising privacy, security and data governance), classify some of their cookies into the categories suggested by ICC [One19, Col18]. However, recent work shows that a large number of cookies in Cookiepedia are categorised as "unknown" [CABM16b]. Multiple studies have used Cookiepedia but completeness has been an issue, with less than 45% of cookie names being recognised [CABM16b, UDHP20, CABM16a], which has impacted the usability of Cookiepedia for cookie categorisation.

To that end, a few earlier studies also looked at tracker categorisation using classification techniques. For example, the timestamp or IP address embedded in cookies has been the basis of unsupervised classification of trackers [GJA+17], while others use application-level traffic logs to automatically detect services running some tracking activity [MTM15a]. In general, even more, studies have attempted to detect privacy leaks via machine learning, from detecting tracking to detecting phishing [JG19, TJH+18, ISZ+20]. In this work, I developed *CookieMonster* which uses a supervised classification approach. *CookieMonster* uses Cookiepedia data as its training data to create a supervised cookie detection framework that is accurate and categorises cookies with very low latency based on features extracted for just cookie names. Furthermore, the Cookiepedia labels are accessible to divide cookies into all four UK ICC categories, rather than a coarse-grained division into tracking and non-tracking cookies as in previous work.

2.4 Privacy Protection Technologies

2.4.1 Privacy legislation

Before 2019, as the new privacy regulations promulgated since May 2018, there is only a handful of GDPR-related measurements and analyses. The initial conclusion from Iordanou *et al.*[ISPL18a] is that tracking flows mostly stay within the EU. In a periodic survey of top 500 sites, [DUL+18] claims that around one-sixth of websites (15.7%) had reorganised privacy policies by May 25, 2018. [UTD+18] investigated cookie synchronisation and show that GDPR cookie consents are insufficient to prevent leakages. Differing from these studies, I examine GDPR over a long duration, using real users' browsing histories and focusing on the GDPR's impact on third-party tracking.

In 2019, Senzing Inc. [Sen19] suggests that around 60% of European companies are not yet prepared for GDPR and 44% of the EU's largest companies are worried about compliance with GDPR. [VEMR+18] studies kids and teenagers' privacy and finds the EU children may be subject to more third-party tracking compared to the US. [LGN18] examines news websites and finds that the UK, in particular, has a high level of tracking. Results such as these corroborate our findings that sites may not be offering a choice, or offering a choice and then not respecting users' choice[2].

Different from studying the behaviour of actual websites is to take an economic, policy or legal perspective. However, even in these fields, it is now being recognised that choice may be difficult for users to deal with, given the complexity of these sites and the technology used [PO16a, HP06a]. Libert, T.[Lib18] developed a tool in 2018 to examine privacy policies of websites to see if all third parties are being disclosed, and finds that privacy policies are extremely complicated, and several third parties are not being disclosed. Figure. 7.6 in the remaining chapters also suggests that in practice users may not make choices that maximise privacy. Therefore, works related to GDPR from then to 2021 focuses on how GDPR impacted the cookie consent notices or "GDPR banners" [HdTS20].

2.4.2 GDPR consent protection

Since May 25, 2018, if companies are involved in collecting, using, processing, and sharing EU/EEA citizen data, it is necessary to comply with the General Data Protection Regulation (GDPR) to implement data protection and corresponding measures. Otherwise, they would face serious financial penalties. And the user's consent is expressly required before processing data; the consent shall be "freely given", "informed", "unambiguous", and "specific" [ICO18].

[2] Example screencast videos for such websites in Top500: https://bit.ly/2GnWrim

Since GDPR came into play on websites for the European Union, European users have been protected online through cookie consent banners. To avoid additional cookie-based positioning losses as much as possible, many websites potentially tend not to place GDPR equivalent protection in non-EU regions [GFdAS21]. While in the EU, the complexity of GDPR consent affects the increased likelihood that users would accept the default GDPR cookie settings by the website [HS19]. Although most users have noticed and are satisfied with the impact of these consent notices, there are still 16% participants in the survey of [Jun18] who thought that web analytics and use of their personal data has not been affected by GDPR.

Under such legal and user requirements, it spawns the consent management platforms (CMPs), in order to assist businesses in informing users of the kind and purpose that data being collected, providing users with the option of opt-in or opt-out of data collection, and notifying other third-party vendors of the user consent. Noted that CMPs support all-or-nothing decisions about the use and provide users with a custom choice of which third parties they want to share data with.

Then, CMPs are popularised for compliant cookie use, enabling website owners to manage service providers and their corresponding cookies, and empowers end-users with giving informed consent for setting cookies. Through a longitudinal measurement of 161 Million Browser crawls, an assessment in 2020 [HWB20] estimated that CMP usage has been doubling annually since the introduction of GDPR in May 2018. Despite the increase in sites that use GDPR CMPs over the past few years, current studies observe that strict compliance with GDPR is still hard to achieve. For example, [NLV+20] demonstrates that websites widely deploy several common dark patterns that make it difficult for users to choose the most privacy-oriented option. This problem is exacerbated by the common usage of CMPs that can quickly distribute such bad practices across the web. Many top CMPs may also use ambiguous consent definitions of cookie usage, such as the notion of tracking under the umbrella of "legitimate interest" and burying it under several layers of obfuscation [PPKM21, JSS+20].

To address this, [MBS20] uses automatic and semi-automatic crawl campaigns from an EU location to identify a variety of suspected GDPR violations among the different CMPs. Some studies are focusing on compliance checks on whether websites comply with GDPR, such as DataCorp [LWED20] and the compliance framework proposed by [Bro19, SRSA19]. Rather than disclosing GDPR issues, this dissertation provides a solution to users that revoke and rewriting GDPR choices from the user side, to decrease the risk resulting from compliance issues. Other studies [UDF+19, DUL+18] have explored the consent and its impact, which provides a good guide for the GDPR notice design. However, neither of them discusses whether the top or specific category websites could gain more trust of users regardless of the banner position and options. Instead, I measure whether the knowledge/attitude of users towards GDPR affects their options, not just focusing on the numerical metrics of notices [VEAWN19].

I first design a survey to bridge a gap left out by earlier works that studied CMP designs from the perspective of visitors of one popular website in Germany [UDF+19]. In contrast, I am more inclined to a comparative grounded user preference analysis for popular cookie consent notice UI designs available widely in the web ecosystem across many websites. I also broaden the respondent pool to users from over 30 countries, finding that this international audience can choose settings that maximise their privacy. Besides, it is also available to elicit from them a direct preference ranking of the different kinds of cookie consent UIs, which can guide websites' future adoption of such designs. The international nature of the survey participants will also make it relevant for other countries that are exploring privacy regulations. Then, based on the statistics on consent preference, I deploy an extension for a user-centred GDPR consent manager in various browsers and accept privacy setting contributions in a peer sourcing mechanism, which fits within the data minimisation in third-party tracking.

In addition, [RSBL] further proposed a "consent recommendation system" based on the survey results of users with active LinkedIn accounts. Whereas this solution is specific to LinkedIn privacy settings, the solution proposed in this dissertation works more widely across the web. There are also a few earlier works that aim to simplify these cookie consent notices and assist users in choosing privacy-preserving settings [HSM21, DUL+18]. However, understanding the ease of use of consent notice interfaces and their efficacy in protecting user privacy remains a relatively unexplored yet important direction.

2.4.3 Plugin-based tools

More and more browser add-ons are currently being launched on the market, which shows that the requirements and rights of users related to online privacy are valued. Related works such as [KNW11, SCM+10, MM12] have been looking at better ways to detect online trackers, including anonymising the *referrer* field in HTTP requests [LHFE15]. Although extensions developed in this dissertation do not directly aim at blocking trackers or attacking them in other ways (e.g.,[LHFE15, KWK+18]), identification of third parties is a paramount first step and a key concern. For this, here has referred to and used strategies, heuristics or third party lists from a number of efforts like ChromeDanger [BCJ+14a], Ghostery [EN16], Brave [FVGJ19], AdReveal [LSW+13a], Adblock [Pap18], Plus [GFLC17a], XRay [LDL+14], TrackAdvisor [LHFE15], and Disconnect [WU16]. Other works focus on advertising [LSW+13a, LHFE15] or service media [SVdB20] alone as well as they do not consider country-specific trackers. Noted that the disambiguation of third parties is required in studies [WU16, KW09] based on authoritative DNS servers, which is also used in this dissertation.

In terms of the ad-blocking performance evaluation, [DLT18] uses inspectors to evaluate the difference of blocking/capturing results caused by different strategies of privacy-enhancing technologies (PETs) or browsers. For example, Ghostery and Disconnect only capture requests but do not modify ad attributions, while uBlock Origin uses filters to change the attributions of ad scripts to block the embedded advertisements. Furthermore, the performance of ad-blockers is examined based on their ability to restrict the raw number of third parties per site and interconnections between first-party websites, respectively. Compared with general ad-blocking tools, tracking-focused browser extensions provide detailed information about trackers to make users more aware of them while being tracked; and are also more effective in blocking online trackers [BLSD+20, MHB+17].

Except that, an increasing number of GDPR related protection extensions have emerged on the market. [RKDK17] provides the introduction of a privacy dashboard and categorises the data subjects. For each disclosed personal information, users would be given the possibility to withdraw consent, and the possibility to request rectification or erasure of the data[3]. The dashboard concept is useful to adopt but they do not follow the study into the internal detection and algorithm. Then, as an improved version of the consent request (CoRe) user interface (UI), [DK19] offers users the choice of the different conditions of consent through the description with the specific consent setting version. [HL21] studies how users connect with the Internet to understand perceptions of privacy/cookie consents on their behaviour towards cookie settings. The results suggest that although the factors that influence interviewees to accept cookies without adjusting cookie settings vary, browsing habits and annoyance are the most influential.

There are various (GDPR-based) privacy protection tools for web users. For example, the user-centric MyWebGuard[PPA+20], deep learning-based Polisis[HFL+18], GDPR evaluation PrivacyCheck [ZB20], etc. However, most prior works do not automatically manage the GDPR settings. The sole exceptions are extensions such as MinimalConsent [Mad20], Consent-o-matic [Uni19] by Aarhus University, Privacy-Cloud Consent Manager [Pri20] and I Don't Care About Cookies [Kla21], each of which detects and mitigates a handful of Consent Management Platforms, mostly through handcrafted values. *CookieCutter* that developed in this dissertation is more extensible than these extensions because of its use of regular expressions and machine learning to detect new GDPR consent cookies, and the use of peer-sourcing to extend

[3] Example site:http://raschke.cc/GDPR-privacy-dashboard/

the database of values to set for maximising user privacy. §8.5.1 presents a more detailed comparison with an example baseline, Minimal Consent.

Chapter 3

Preliminary works: Data collection

Summary

This chapter mainly introduces the methodology of data collection we used for third-party tracking and measurement in-the-wild, as well as the collected datasets concerning the browsing histories and consent details. In addition, the research ethics presents that the informed consent we intend to collect from real users and the conduct of protecting their data anonymity, which respectively approved by King's College London research ethics committee (No. MRS-1718-6539) and University of Surrey ethics board (No. 514292-514283-64690499).

3.1 Data collection using browser extension

The data collection methodology in this study is mainly based on the participation of real users in-the-wild. Since our study is a human-centred measurement, there are normally two kinds of data collection methodologies.

The first one is passive data collection, which is basically takes place without participants' awareness, i.e., "non-active participation." That is to say, the data collection could be completed without initiative actions of participants towards the research. For example, both smartphone usage collection via mobile apps [KSA+19] and clinical data from patients [MSH+19] belong to the ongoing passive data collection. The other is Observational research, which requires participants to actively carry out particular activities in dedicated environments. And sometimes,the researcher needs to be involved.

As we all know, encouraging large numbers of people to choose to participate in a particular study is challenging, especially where it feels like we are collecting (potentially sensitive) data from them. From this perspective, passive data collection is relatively easier to recruit research participants for as there requires minimal effort. Compared with observational research, passive data collection tends to be on a larger scale, but with the less difficulty of research recruitment.

Since browser extensions are generally used to strengthen the control in browsing behaviours of online users or provide additional network service functions [Goo21], it is beneficial to attracts users with different demands, expertise and privacy stance. In such a case, by constructing a browser extension (passive data collection), the range of samples or targets that available for selection is much larger than other survey or research methods.

3.1.1 Introduction to Thunderbeam

Thunderbeam is a Chrome extension developed by us for research purposes.[1] It offers an alternative version of Lightbeam [Moz12] for Chrome users. In addition to considering that many of our real-data contributors prefer to use Google Chrome, and Lightbeam in Firefox has gradually becomes the first choice for privacy-related add-ons. Therefore, We extend the Firefox extension to support its work on Google Chrome, to unobtrusively adapt to the browsing habits of our users. Making the extension work on Google Chrome involved tackling several challenges, which we outline below.

Lightbeam is a branched add-on of Collusion [Too12], which developed by Mozilla in 2012. And another branch is a Chrome add-on named Disconnect [Dis13]. Light-

[1] Thunderbeam-Lightbeam for Chrome: http://tiny.cc/thunderbeam-chrome-plugin

beam is at an advantage because Disconnect in Chrome can only record trackers in a non-contextual manner (i.e., by viewing the website separately). The UI of Disconnect is only able to offer users the visualisation graph of third-party relation in the current tab, as it does not support cross-site tracking of third parties. As opposed to Disconnect, Lightbeam in Firefox could draw the tracing graph across the user's browsing history, providing support to track third-party networks across sites.

The reason is the browser compatibility of a particular API in Chrome, which results in the limited functionality of Disconnect on the Chrome browser. Thus, there is no direct mechanism in Chrome to capture and match the correspondence between first-party and third-party requests. Therefore, Chrome fails to tie the third-party requests to the first parties that initiate the loading of those third parties, maintaining the same mechanism with Collusion and Lightbeam. Specifically, Lightbeam applies the WebRequest API in Firefox contains a property called `webextensions.api.webRequest.onBeforeRequest.details.originUrl`, when a HTTP request is made to any website. In a third-party HTTP request on Firefox, the `originUrl` gives the details of the first party that initiated the request. Unfortunately, `originUrl` is not supported on Google Chrome, that is to say, Disconnect does not track third parties across the browsing histories is due to the lack of direct mechanism in Chrome like Firefox that Lightbeam uses to connect the third-party requests with the first parties.

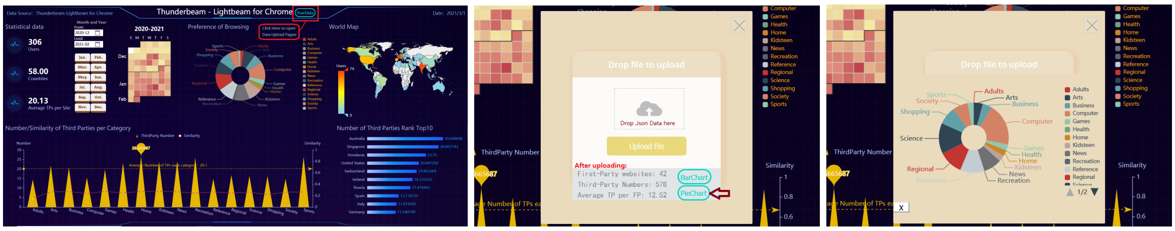
This dissertation resolves this limitation by noting that the the `tabId` field of `WebRequest` API that Chrome supports, which identifies the index of tab which made the request. Therefore, I build and maintain a table containing all open tabs, and use the URL loaded by the tabs to obtain first-party information. Then, I utilise the correspondence between *tabIDs* to classify the first-party attribution of each third-party request.

To use tabId to build the relationship between first/third parties, chrome.webRequest.onResponseStarted.addListener is first used to capture all request URLs and the corresponding tabIds of each request. Although originUrl of the request is not accessible, a match-up table could be completed between the first-party and third-party requests by aggregating the same tabId to achieve the same effect. Once entering a first-party website that ever visited before (after installation of the plugin), chrome.tabs.getSelected is used to capture the first-party domain by retrieving the value of tabId. Then, placing the host name of the first party url in the match-up table to correspond to the specific tabId. Therefore, other remaining requests could compare the value of their tabId they located at with the index of the match-up table to identify the third-party tab position . In the meanwhile, it determines whether the request belongs to the first party or third party through the comparison between the provider and the first party provider that retrieved by tabId.

However, there is another approach (building a nested loop) to achieve similar functionality on the currently visited website. Briefly speaking, the "outer loop" is to execute the JS injection to the client-side console with performance.getEntriesByType() to obtain the list of all resource url that placed in the current website. For each iteration of urls, the "inner loop" executes the cookies.getAll() function once for that particular domain, receiving all cookies set with the url request. In this way, the extension needs to perform a nested loop for each visited website, thus the complexity of identifying the first-party and third-party objects reaches O(n^2). Furthermore, chrome.tabs.executeScript costs additional 1 millisecond. The two main advantages of this approach are as follows:

Advantage 1. Complexity: Compared with the nested loop, the complexity of the proposed method with the match-up table is lower, which is O(n). Since the extension would continue to be processed in the back-end each time the user accesses or refreshes the tab, it is necessary to ensure that the complexity and process cost of the function are both as low as possible. Otherwise, it would not only fail to improve the control of the browser, but also reduce the user's browsing experience due to the increase in latency.

Advantage 2. Ongoing tracing of inactive tabs: Apart from the complexity, another benefit is the ability to continuously tracing all tabs regardless of the changes of active tabs. Although the focus moved to the other tabs, the match-up table could still



```
First-Party websites: 42
Third-Party Numbers: 578
Average TP per FP: 12.52
```

(a) Screenshot of the data visualisation and analysis website

(b) Self-Service data visualisation

(c) Bar chart of data visualisation Self-Service

Figure 3.1: Users are provided with an analysis suit to asses the privacy implications behind their browsing habits w.r.t. the third-party ecosystem.

maintain receiving requests from all tabs. It does not stop the tracking towards third-party requests in the remaining tabs. For example, the nested loop would inject the client-side console once, which returns the number of resources loaded at the moment of injection occurs. It is hard to track the continuous state for all tabs in the window, so the `webRequest` API with a match-up table is the better choice.

To preserve the anonymity of our users, we use a randomly generated UUID (Universally Unique Identifier) to identify them. We discuss the ethical implications our research in Section 3.3. The length of the UUID is 128 bits, of which 122 bits are random bits and 6 bits are reserved. It makes use of Universally Unique Identifiers (UUID) version 4 variant 1 as defined in RFC 4122 [Aro18] to describe how we prevent potential leaks by using UUID instead of database ID, in URL and API. Meanwhile, the identifier length also ensures a very low probability of collusion.

3.2 Datasets

This section mainly includes three key points: browsing history data (first-/third-party data), cookie settings of GDPR consent banners (GDPR consent cookies) and cookie classification statistics (four categories of cookies)

3.2.1 Tracking dictionary

This dissertation describes the method of using §1.2 to assess the privacy loss caused by potential data sharing based on cookie synchronisation [PKM19], which is another "data sharing tunnel" mechanism between different third-party entities.

Optimised third-party database. The third-party database consists of different existed online tracker dictionaries, and Table 3.1 provides the details **such as the number of each category, as well as in which literature they have been used and studied**. We classify third parties as follows according to their functions.

1. **Advertising Third Parties**. It is usually provided by the advertisement server used by the demander of the advertisement service or the advertiser. It includes two subcategories: i) advertisements or services from other coorperated first-party entities, and ii) advertisements from third-party networks (for example, *doubleclick.net*, *adkernel .com* or *webspectator.com*).

2. **Analysis Third Parties**. It is mainly provided by a website analysis company to evaluate the operation performance of the website and user experience "feedback‘‘ without notice. (For example, *google-analysis.com*, *hupso.com*, and *audienceinsights.net*.)

3. **Essential Third Parties**. Refers to the domains that are required to support the basic operating functions of the website, such as secure login, cloud storage of website resources, etc. (for example, *bootcss.com*, *squixa.net* and *commanders-act.com*.)

4. **Malware Third Parties**. This category generally presents third-party domains that potentially cause serious privacy risk, including adware, viruses, and potential ransomware. (Examples include: *msecnd.net*, *imrworldwide.com* and *securestudies.com*.)

5. **Optimization Third Parties**. These domains provide optimization related services to obtain the increase in productivity and experience of user, such as

support for higher speed, automation, and interactive marketing, etc. Realize the more and better in control of the performance of both first and third party assets.(E.g., *yieldoptimizer.com*, *maxymiser.net*, *bzgint.com*.)

6. **Redirect Third Parties**. Domains that use HTTP redirection to “share” the user traffic from one domain to another, offering an alternative to existing third parties that have been placed. Some of them are URL shorteners and third-party payment services. Please note that this category would change the URL destination without any user interaction or notification. (E.g., *redirectingat.com*, *tinyurl.com*, and *clickredirection.com*.)

7. **Social Third Parties**. This category usually consists of third parties from a number of social platforms, most of which use social media widgets to collect user browsing behaviors on other websites for social media companies. The recorded user profile is then used to improve the advertising positioning that shown on its social media. (For example, *metabroadcast.com*, *facebook.com*, and *twitter.com*.)

8. **Tracking Third Parties**. This category includes all forms of tracking domains through embedded technology, web bugs, and providing customer’s personally identifiable information (PII) data to third-party providers. (For example, *otracking.com*, *tctm.co*, and *zenfs.com*.)

As the eight third-party categories listed in Table 3.1, we merge different open source third-party databases and confirm a unified nomenclature for these third parties. The different lists mostly agree on the category of a particular third party (which appears only once). Where there is a disagreement (for example, a particular third party is classified as social on one list and malware on another), we use a majority rule to eliminate the ambiguity. To be more specifically, Facebook supports various categories like social, tracking, and advertising, while Google occasionally stands by either advertising or analytics. Therefore, we need to calculate third parties in all relevant categories, instead of overruling solely relied on a single count.

TP Categories	#Num	Source Database
Advertising	9110	Disconnect(1588) [Dis], Webpage Toaster(1013) [Dai], EasyList(7580) [Eas], pgl(2816) [Low], **OurList** (551)
Analysis	552	Disconnect(275), Webpage Toaster (308)
Essential	965	Disconnect(530), Webpage Toaster(515), **OurList** (47)
Malware	68	ZeuS Tracker [Zeu], MalwareTips [Mal], **OurList** (68)
Optimization	582	Webpage Toaster(394), **OurList** (188)
Redirect	31	MalwareTips, **OurList** (31)
Social	157	Disconnect(59), Webpage Toaster(73), **OurList** (68)
Tracking	3128	Webpage Toaster (74), EasyList (2088), Better(604) [Bet], WhoTracks.me (1130) [Who], **OurList** (732)
Total	11,232	Of this, 1685 were manually checked and added to "OurList", *strengthening* the "Chinese tracker list" **(i.e., the trackers placed by Chinese vendors)**.

Table 3.1: Number of domains in each third-party category observed in our users, as identified by eight different third-party databases. For each source and category we list in parenthesis the number of third parties.

To get around the limitations of current Web tracker lists (which are especially poor for Chinese websites [SRDK+19b]), we also manually classify TPDs which are not in any of the known lists. Our manual categorization involves visiting the website of the third-party vendors to check the "Home" and "About" pages, checking the JavaScript used by the third parties, and querying Open Source Intelligence for vulnerable/malicious indicators. Overall, we identify a significant number of third-party domains (1,685). We refer to this manual annotation as "Ourlist" in Table 3.1.

3.2.2 First-/Third-party browsing history

According to the tracking mechanism, when the user is active in browsing websites, we could capture the relationship between the first-party website and an array of third-party trackers.

This study develops two new datasets to explore the third-party ecosystem: one based on real-world browser histories and another one based on the most popular

Alexa websites. The first dataset looks at three user groups (UK users in UK, Chinese users in China, and Chinese users in UK) and I compare the structure of third party networks across groups. This is complemented by users from two other countries, Australia and USA. The China and UK data is longitudinal for over a year, and the Australia and US data represents nearly one month of activity. This dataset is used to answer RQ2 and RQ3 (see §4.4).

Using Thunderbeam add-on, there are three periods of real-user data collection and one additional Alexa-based dataset considered in this chapter. The first one of real users is obtained after collecting anonymised browsing histories through the new Thunderbeam add-on from 16 different users weekly between January 5, 2018 and January 2019 — 9 users in the UK and 7 users in China (CN). Here, 3 of the 9 UK users are of Chinese origin and tend to also visit Chinese websites from the UK. I term these users as CN-UK, and they offer a unique perspective on the tracking done by Chinese websites to users out of China. Then, more users from Australia (AU) and United States (US) enrolled in this study to determine the third-party ecosystem similarity between Alexa and real-user study.

Table 3.2 shows an overview of the first-phase data collected from these users. In total, 15,323 unique third party domains are gathered from the browsing logs of research participants, between Jan 5, 2018 when the collection started, and Jan 1, 2019, the last date reported in this chapter. Of these, 1,685 sites were manually identified as described above, yielding a 15% improvement over the union of previously known lists. The extension could successfully categorise nearly 90% of third-party domains for the UK users and 70% of third-party domains for Chinese users.

The last anonymised real-user dataset are from 2484 users in 102 countries (among 9506 **installers in the Chrome Web Store**), being collected weekly between December 2020 and August 2021. But the total number of users listed by country is 2537, which results from that the same set of users (identified with UUIDs) are contributing data from different countries. The demographics of the users that have participated in this large-range study in Table 3.3 and 3.4. Although there are representatives in all

User Group	**Records of FP**	**Records of TP**
UK users	8416	113,003
(include CN-UK users)	(2680)	(36,209)
CN users	6144	74,313
US users	392	4450
AU users	104	820
Total	15,323	192,586

Table 3.2: First-party (TP) and Third-party (TP) data collected from participants.

continents, most of users are distributed in Europe and North America (especially the United States. **According to the status of user data collection,** users in these regions lean more privacy-conscious. Besides, it is observed that countries which have enacted privacy protection legislation tend to contribute more users. For example, GDPR for Europe, C, etc.CPA for California (US), LGPD for Brazil, PDP Bill for India.

And the second part of datasets relied on `Alexa.com` is obtained using a controlled experiment. In particular, Selenium is paired with Thunderbeam plugin to leverage top500 sites from Alexa. **Alexa lists have been used extensively in many research projects and are proven to be one of the most well-known resources for relatively accurate and easily accessible website rankings.** Noted that Alexa has stopped the support of providing lists of top web sites by category [Ale20b], thus I use the version of a web archive that screenshoted in July 21, 2020 [Ale20a].

Table 3.3: Data contributed by 2416 users in 45 active countries from Dec. 2020 to Aug. 2021.. The active country refers to the country with more than 5 users.

Continent	Country	#Users
Africa	Egypt	119
	South Africa	22
	Algeria	8
	Morocco	7
	Tunisia	6
Asia	India	159
	Turkey	30
	Saudi Arabia	26
	Emirates	22
	China	21
	Japan	18
	Bangladesh	13
	Singapore	12
	Indonesia	8
	Sri Lanka	8
Europe	France	289
	Germany	251
	Spain	170
	UK	120
	Switzerland	95
	Italy	82
	Austria	60
	Netherlands	60
	Denmark	41
	Belgium	34
	Hungary	30
	Russia	26
	Czech	19
	Norway	19
	Poland	18
	Sweden	17
	Croatia	11
	Ukraine	11
	Romania	9
	Portugal	8
	Finland	6
North America	US	314
	Canada	74
	Mexico	27
Oceania	Australia	24
South America	Colombia	50
	Brazil	29
	Chile	20
	Peru	14
	Ecuador	9

Table 3.4: 121 users in 57 inactive countries.

Continent	Country	#Users
Africa	Togo	5
	Senegal	3
	Reunion	2
	Congo	1
	Ghana	1
	Libya	1
	Nigeria	1
	Tanzania	1
	Uganda	1
Asia	Israel	5
	Jordan	5
	Philippines	5
	Malaysia	4
	Cyprus	3
	Iran	3
	Lebanon	3
	Thailand	3
	Iraq	2
	Kazakhstan	2
	Kuwait	2
	Qatar	2
	Viet Nam	2
	Bahrain	1
	Georgia	1
	Kyrgyzstan	1
	Maldives	1
	Nepal	1
	Oman	1
	Syria	1
	Turkmenistan	1
Europe	Greece	5
	Estonia	4
	Luxembourg	4
	Serbia	3
	Belarus	2
	Ireland	2
	Slovakia	2
	Slovenia	2
	Andorra	1
	Bosnia	1
	Bulgaria	1
	Iceland	1
	Malta	1
	Monaco	1
North America	Costa Rica	2
	El Salvador	2
	Honduras	2
	Barbados	1
	Guadeloupe	1
	Guatemala	1
Oceania	New Zealand	4
	Wallis	1
South America	Argentina	3
	Paraguay	3
	Uruguay	3
	Bolivia	2
	Venezuela	1

UK	China	Australia	US
{college website}	tmall.com	aliexpress.com	google.com
google.com	baidu.com	google.com	{college website}
microsoft.com	weibo.com	renren.com	linkedin.com
wordpress.com	taobao.com	ebay.com.au	wix.com
stackexchange.com	qq.com	youtube.com	pinterest.com

Table 3.5: Top5 first-party domains in the real-user database. **Noted that a large user base in the UK and US is from universities, this may be an indication that the study might have some bias on personnel demographics due to the need for a certain knowledge and willingness to download the browser extension.**

Then, the browsing habits of users are understood by measuring the average number of visits per user per week. In particular, I measure the number of sub-pages (URLs) visited and the unique number of the second-level domains users connect to Figure 3.2 shows that the data is characterised by a set of relatively active users, which visit tens of unique domains. In particular, these users visit over 100 web pages per week. Within each second-level domain site, users visit on average around three pages. Figure 3.2 defines a dashed lines to highlight the average browsing habits of half of the users. It indicates that half of the users visit at least 50 unique first-party domains and 65 unique web pages each week. When comparing the distribution of the number of unique FP domains with the number of FP pages, it can be seen that the trend in both distributions is uniform. The bulk of the users browse like the average user, although naturally there are some fragmented users that browse deviate from the average.

Table 3.5 summarises the most frequently visited first-party domains per country. A key distinction on the way I compute the frequency of visited domains with respect to `Alexa.com` is that I look at second-level domain names. This way I better capture the organization that registers a domain name, making the data more analytically usable.

Second, I contextualise findings by looking at the prevalence of third parties in some of the most popular sites (according to `Alexa`) in an automated fashion. The second dataset has been obtained using a controlled experiment. In particular, I have

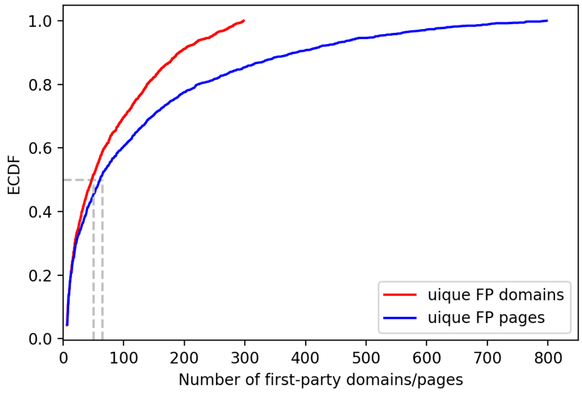


Figure 3.2: CDF of the number of unique first-party (FP) web domains/pages visited weekly overall.

used Selenium instrumented with the Lightbeam plugin to crawl popular sites from `Alexa.com`.[2] For the purpose of this chapter, I mainly leverage the `Alexa top2000` global websites,[3] the `top500` categorized sites in 16 categories **(8,000 sites in total)**,[4] and `top500` national sites from each country[5].

The rationale behind this experiment is to understand how representative the four-country dataset is Figure 3.3 depicts the empirical cumulative distribution function of the number of third parties placed in websites that accessed from four countries, i.e., UK, China, US, Australia. Each country has two lines representing data collected from the real-user study and the Alexa topsites database to compare authoritative data sets in the same country/region with real-world data sets. Here I pick up four representative countries, and the result shows that the distribution of the numbers of third parties per domain in the user dataset **mostly** matches that of the topsites dataset for each country. This gives the confidence that the data obtained from the real-user group maps with third-party privacy behaviours of each of the most populations. **However, the ranked websites might also provide an upper bound on the amounts of tracking. Although the real users with browsing habits are consistent with the**

[2] https://www.alexa.com/topsites, which provides researchers with top websites across countries and categories

[3] Global Alexa top: aws.amazon.com/cn/alexa-top-sites/

[4] Alexa by category: www.alexa.com/topsites/category

[5] Alexa by country: www.alexa.com/topsites/countries

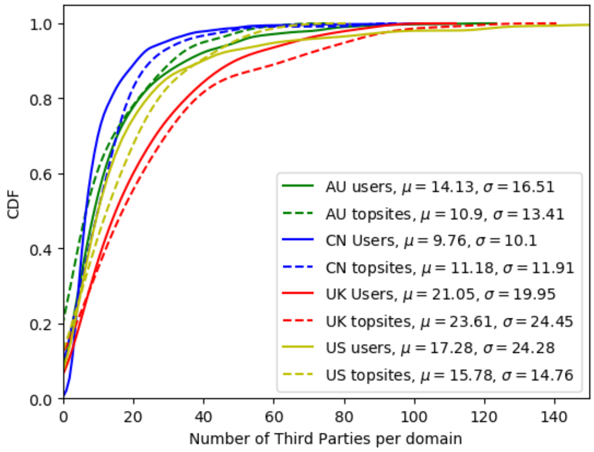


Figure 3.3: Number of third-party sites per domain in User Group data (*real users*) follows a similar distribution to the numbers of third parties on `Alexa top500` (*topsites*) in each country. *UK users also have more third parties per domain than CN (China) and AU (Australia). CN-UK users visit both CN and UK websites, and also without the Great Firewall of China, and thus experience an intermediate number of third parties between purely CN and purely UK users. Surprisingly, both data from US users and US topsites show the fewer third-party number than UK.*

general tracking trend, there still exists specialist-specialist or location-specialist difference in terms of tracking risk. These observations answer RQ1 and RQ4 in §4.3. The assessment of the `Alexa` top websites uses Selenium in non-headless mode to simulate user activity, which allows the collection of connections between first parties and all dynamic third parties on an active browser. Then, I repeat the test 25 times per day over 7 days to confirm that change in numbers of dynamic TPs is less than 1% (c.f. §4.3.3).

Figure 3.3 also highlights some interesting results: the differences in the number of third parties found across countries suggests that UK users have the most third parties per domain than any other countries. Since CN-UK users visit first-party sites from both CN and UK websites without the restriction of Great Firewall in China, thus it locates at an intermediate number of third parties between purely CN and purely UK users. Namely, the number of third-party providers hitting CN-UK users is higher than for CN users alone — yet, not as high as UK users because many of the websites they visit are Chinese websites, with lower levels of tracking as discussed later.

Furthermore, the number of third parties in US is fewer than UK's, which implies that the popularity of third-party providers is in general higher among UK-based users than US's, with demographics and users' browsing habits(not just the location) playing a crucial role. Related works forementioned have provided a comparisons between region-specific third parties [FHUM14b]. However, their analysis classifies regions by language, resulting in differences between US, Australia and UK that do not emerge.

This also accords with the earlier observations in [HdTS20], but the larger size of cohort enhances the distinction between real-user and Alexa datasets. But the distribution of the numbers of third parties per domain in the user dataset still roughly matches that of the topsites dataset for each country. This gives the confidence that the data obtained from the user group maps with behaviours of each of the populations.

I examine where trackers are most effective based on this finding, for example, which categories of sites are they most prevalent in; how well they can track individual users as well as cohorts of users. This reveals differences across sites and categories that are common in both countries (§4.3), and differences between tracking across countries (§4.4).

3.2.3 GDPR consents

The term "GDPR consent cookies" is derived from General Data Protection Regulation (GDPR) and refers to the user's acceptance of the cookie set in the GDPR notification and the corresponding expiration date of choices. In some cases, the website considers using an additional cookie to determine whether the identified user has made valid operations for the GDPR consent box. For example, www.theguardian.com create two GDPR consent cookies for European users to selectively switch off unnecessary cookies, using *_sp_v1_consent* and *consentUUID* to confirm that the cookie consent box has been operated by the particular user.

Due to the different cookie acceptance into the website (e.g., rejection for all), restrictions on certain cookie categories could effectively reduce the third-party exposure of users. However, it still reflects various shortcomings, when GDPR is applied to websites in scope of Europe that need to collect user information for commercial operation.

When studying on GDPR consent in websites, we can mainly categorise websites into three types: cookie notice with customised options, cookie notice but no customised options and no cookie notice. In terms of GDPR consents, our results are in the basis of two datasets.

The first is the distribution of GDPR consent box categories in popular websites defined by Alexa, indicating the extent GDPR applies to high-traffic websites. The other one is the consent cookie set in-the-wild collected from the Cookie Consent Management (§**8**). It involves 96 users located in 25 countries. Benefiting from real users, we could estimate the long-term momentum of the application of the GDPR consent management platform (CMP) on the website, e.g., OneTrust, TrustArc, etc.

3.2.4 Determine cookie consent notices

Cookie consent notices work by actually setting a small set of additional cookies whose values record users' preferences on what data they permit the websites to collect, whether for tracking or other purposes. I ask a simple and straightforward question: If I am able to *programmatically* set the right values, does this prevent tracking?

To test this, I exhaustively examined all the Alexa top 100 websites in the UK (rankings as of Jan., 2021). Interestingly, only 55 these websites present users with cookie consent banners, although these sites collectively use the top 8 CMPs, representing over *65%* of the CMP market. Determining whether the other 45 may be in violation of GDPR or may not be collecting Personally Identifiable Information (PII) is outside the scope of the study[6]. Instead, I focus on the 55 websites which do ask users for consent to collect their data and manually record the names and

[6] although I note that 55 is a slight improvement over a previous 2019 study that reported only 42 of the Alexa UK Top100 sites offered users a cookie consent notice with choices [HS19]

Cookie type	max	mean	median	min
Advertising	1	0.2	0	0
Analytics	3	0.6	0	0

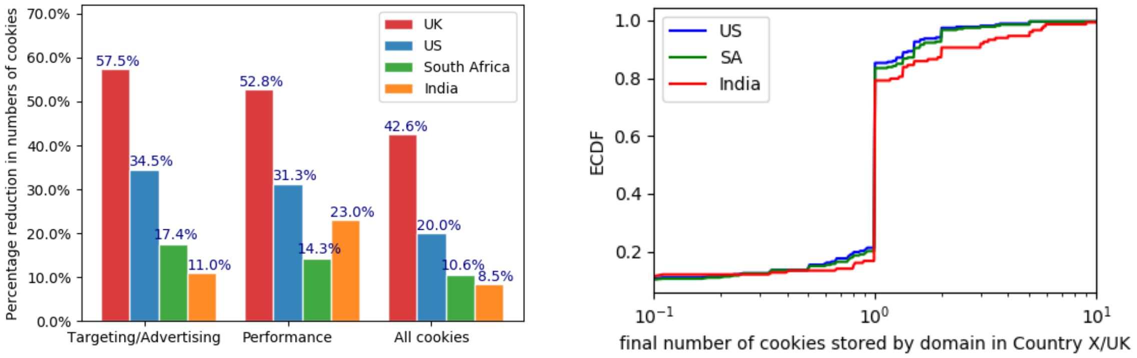
Table 3.6: Difference in numbers of advertising and analytics cookies when a user manually chooses options via a cookie consent notice vs. when GDPR consent cookies and their values are pre-populated in their browser's cookie jars.

values of the cookies which are set when the most private CMP options are chosen by the user. I call these cookies as *GDPR Consent Cookies*.

I then create a new browser profile and *pre-populate* it with the GDPR Consent Cookies previously identified. Next, I visit the 55 websites again and confirm that the user is not presented with a CMP banner in any of the websites. I compare the numbers of (non-GDPR Consent) Cookies set by these websites (for tracking, analytics, etc.) when a user manually chooses the most private option by clicking on the cookie consent banner vs. the number of non-GDPR Consent Cookies set when the GDPR Consent Cookies of these websites are pre-populated for the user. I find that both methods yield similar numbers of cookies (Table 8.1 shows a median difference of 0 for all categories of cookies and a maximum difference of 3 analytics-related cookies), which confirms that *programmatically setting cookies not only saves users from having to choose the right GDPR consent options on each website but also achieves similar effects in decreasing data collected about the user* (solving problem 1).

Next, I check whether the technique of pre-populating GDPR consent cookies in users' browsers works *outside EU locations*. Article 3 of GDPR[7] stipulates that the regulation applies to "applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not." Thus, I expect that websites operated from or controlled by EU-based establishments should obtain consent from all users globally. Similarly, Article 3 also states that establishments outside of the EU must respect GDPR for data subjects in the EU

[7] https://gdpr-info.eu/art-3-gdpr/



(a) Percentage reduction in cookie numbers in four countries after applying the proposed approach

(b) CDF of number of cookies left after applying the proposed approach, for different websites accessed from country X, shown as a fraction of the number of cookies for the same website accessed from UK

Figure 3.4: Cookies on 55 Alexa Top websites when accessed from UK, USA, India and South Africa.

for "monitoring of their behaviour as far as their behaviour takes place within the Union", thus I would expect (and do observe) that global websites to also deploy cookie consent functionality from a vantage point within the GDPR jurisdiction region. From a location outside this jurisdiction, I observe that the websites load the consent management scripts and libraries but they detect the user's location and do not show the cookie consent notices. However, I observe that manually setting GDPR consent cookies still triggers protection against tracking and decreases the numbers of cookies even in regions outside of GDPR jurisdiction.

To quantify this systematically, I use a VPN solution with four different end points — UK, USA, India and South Africa (locations were chosen to be populous countries representative of the different continents of the world; but where reliable exit points were available for the chosen VPN solution. For Asia, India was chosen instead of China because many websites are not accessible behind the Great Firewall of China. The VPN did not support exit points in Australia or South America). I then visit the same 55 Alexa top sites as above, first without setting any GDPR Consent cookies and then after setting GDPR consent cookies. Figure 8.2(a) shows the percentage reduction in cookies in each country, showing that *approach*

***can have an effect globally, even outside of the EU*. Although it appears that some countries have a greater amount of reduction, this is simply because of the larger number of trackers initially before the protection is applied. This is corroborated by previous studies for example [HdTS20] which shows that UK has more trackers than the US, which means that the reduction in the UK after removing all trackers would be proportionally more than in the USA. Similarly [SM19b] shows that UK has more trackers than the USA which in turn has more trackers than South Africa, followed finally by India. This rank order corresponds to the different percentage reductions I see in Figure 8.2(a). Figure 8.2(b) confirms that after applying the approach, most websites have the same number of (essential) cookies left in the three countries outside the jurisdiction of GDPR (*i.e.*, USA, India and South Africa) as from UK vantage points (which is in the jursidiction of GDPR). This suggests that simply setting GDPR consent cookies could afford users in countries around the world similar privacy protection as provided by GDPR in the EU (Problem 3).**

Finally, given that all the above evidence suggests that GDPR Consent Cookies appear to govern website tracking behaviours, I check whether removing the GDPR consent cookies which were previously set offers a way for users to revisit or change their previously set cookie consents. When I manually remove the identified GDPR consent cookies and refresh the web page, I notice that the GDPR consent banner pops up again, allowing users to set a different value for their consent (Problem 2).

3.2.5 Cookie Classification

For this section, cookies cookies are divided into four categories according to the classification pattern proposed by UK International Chamber of Commerce —— Strictly Necessary, Performance, Functionality and Targeting/Advertising cookies (detailed explanation would be given in §7). Cookiepedia [One20] is a database containing more than 37 million pre-categorised cookies, designing to keep a knowledge base of all cookies on the web. It is busy categorising cookies according to widely used ICC cookie categories and adding more cookies collected using tools from OneTrust.

To examine the completeness of Cookiepedia, it is necessary to create the ground truth dataset for labeling cookie names with Alexa top20k (identified cookies in Cookiepedia). The first four categories in Table 3.7 are consistent with the UK ICC categories. If the cookie name does not exist in its database, Cookiepedia returns “nonexistent” ; if the cookie name exists in the database but has not yet been classified, it returns “unknown” . 78.83% of cookies from Alexa top20k website are either unknown or nonexistent.

Cookie Category	# cookies	% cookies
Strictly Necessary cookies	3,071	5.61
Functionality cookies	1,102	2.01
Performance cookies	3,025	5.53
Targeting/Advertising cookies	4,380	8.01
Unknown cookies	19,007	34.75
Nonexistent cookies	24,108	44.08
Unknown+Nonexistent	43115	78.83
Total	54,694	100

Table 3.7: Creating the ground truth of cookie classification model from the classification results of Alexa global top20k websites by Cookiepedia.

Based on the classification result, it could be observed that only 54,694 unique cookie names placed in Alexa top20k websites are validly recognised in Cookiepedia. Therefore, due to the incompleteness of Cookiepedia, we need to design a classification model named *CookieMonster* in §6.

3.3 Research Ethics

I ensure that the research follows ethical principles by applying the guidelines from The Belmont Report [Bea08], controlling the process of data collection and ensuring the privacy of real users. First, the study does not request any sensitive information concerning personally identifiable data such as name, gender, age, etc. nor do we collect information that could be used to identify users like IP addresses. For those who consent, no personal data belonging to the users will be used (and data is hashed to prevent re-identification); any data collected will be used only for the purposes of non-commercial research, and all data will be deleted after the research project completes.

Out of all the data that could be extracted from the browsing history, only first-party and third-party domains would be retained, stored with the the users' consents to extract and study this information. It means that I do not collect URL parameters, which may include username, passwords or other identifiable information.

The consent is to guarantee users proceed in steps: 1) understand the purpose of this study; 2) understand and access their rights; and 3) notice the toggle switch to easily opt-in/out from the data collection, ensuring the anonymised browsing history that they are willing to share with us is only collected. The information sheet and consent form provided to participants of the real-user data collection have been reviewed by our Institutional Review Board.[8] Furthermore, the released Chrome extension provides a Privacy Policy with information about user's rights, including withdrawal.

[8] The consent process has been vetted by the King's College London Research Ethics Committee. The criteria for approval can be found here: https://bit.ly/2XHiwT8

Chapter 4

Third Parties from Multi-country User Experience

> *Privacy means people know what they're signing up for, in plain language, and repeatedly. I believe people are smart. Some people want to share more than other people do. Ask them.*
>
> —— Steve Jobs, Entrepreneur

Summary

In this chapter, the key objective is to understand how the third-party ecosystem is developing in the different countries of the real-user study. It is interesting to investigate how wide a view a given third-party tracker may have, of an individual user's browsing history over a period of time, and of the collective browsing histories of a cohort of users in each of these countries. I study this by utilising two complementary approaches: the first uses lists of the most popular websites per country, as determined by Alexa.com. The second approach is based on the real browsing histories of a cohort of users in these countries. And the larger continuous user data collection spans over a year. Some universal patterns are seen, such as more third

parties on more popular websites, and a specialisation of trackers with presence in some categories of websites but not others.

However, the study reveals several unexpected country-specific patterns: China has a home-grown ecosystem of third-party operators in contrast with the UK, whose trackers are dominated by players hosted in the US. UK trackers are more location sensitive than Chinese trackers. One important consequence of these is that users in China are tracked lesser than users in the UK. The unique access to the browsing patterns of a panel of users provides a realistic insight into third party exposure, and suggests that studies which rely solely on `Alexa` top ranked websites may be over estimating the power of third parties, since real users also access several niche interest sites with lesser numbers of many kinds of third parties, especially advertisers.

4.1 Introduction

Web advertising has evolved considerably over the last decade. Publishers and advertisers currently leverage Web technology to track users' browsing histories and make advertising even more targeted, and therefore more profitable. This is supported by many of the most popular websites that willingly embed this technology into their sites to monetise the content they host. This technology basically allows publishers to obtain a unique identifier of the visitor of a site, which is then used to match the user across other websites. Although there are many ways in which a tracker can associate unique identifiers to visitors, current efforts are largely based on the DART (Dynamic Advertising Reporting and Targeting) initiative launched by DoubleClick [RF09]. Here, unique third-party cookies are left in the browser of the user when she visits a website with a tracker embedded. With the scaled size of the advertisement industry, this poses a risk to the privacy of the users and leads their browsing history being shared in some shape or form with publishers and advertisers.

Related works in the area have recently looked at this problem and they have provided a good understanding on how the tracking technology works and the underlying privacy issues [RKW12, LHFE15, MTM16, HLM18], including mobile tracking [RNVR⁺18]. However, their analysis only look at a slice of the problem: either because they look at a specialised third-party network [ISPL18a], or because they look at the problem from a holistic perspective without considering users' browsing patterns [EN16]. For instance, authors in [LHFE15] look at advertisements alone, [RKW12] and [EN16] do not look at the population segmentation, and [HLM18] does not quantify how third-party categories change over time. A central aspect of understanding the tracking ecosystem is characterising the different trackers users came across. This is a challenging process as the third-party ecosystem is complex, highly dynamic, and — in some cases — localised [SVdB20]. The study differs from other works in the scope of the analysis. Here, I consider general-purpose third-party domains with a fine-grained categorisation. I also consider, as a key distinction, targeted population segments (e.g., Chinese users) in different locations (i.e., domestic users vs. users abroad). Not only does the project provide a comparison of third parties across countries and categories, but it also provides insights into the causes of differences based on a broader data collection.

To better understand the magnitude of the tracking problem, I present the following main contributions. First, I build technology that can capture to what extent third-party trackers are profiling users as they browse the Web.It is addressed by extending a popular Firefox extension, called Lightbeam, in two directions: i) enabling the support of fine-grained cookie logging and porting it to Chrome, and ii) integrating an automated browsing system into it. That extension is available in the Chrome Store as the "Thunderbeam-Lightbeam for Chrome" plugin[1] and has seen 9,506 installs (as of October 9, 2021).Second, an improved categorisation (15% improvement) of the type of third-party providers is provided by employing a number of heuristics and using several online resources. I freely make available the resulting "Tracking the Trackers categorisation list"[2]. Finally, I study the interplay between user location and the overall number of third parties observed using a twofold approach: with an

[1] https://tiny.cc/lightbeam-chrome-plugin
[2] https://tiny.cc/tracking-trackers-list

automated controlled experiment and a user study. In particular, I look at third-party domains in five different population segments: Australia (AU) users, United States (US) users, United Kingdom (UK) users in UK, Chinese domestic users (CN), and Chinese users located in UK (CN-UK).

4.2 Research Questions

This chapter aims at answering the following research questions, which I present together with main findings:

RQ1: Is the number of trackers per site affected by the popularity of the website, as well as its category?

§4.3 demonstrates specialisation of third parties across categories. Thus, third party actors are more easily able to track individual users (who fit their specialisation areas) across time, than a diverse cohort of users simultaneously.

RQ2: Are there country-specific third parties?

§4.4 finds specialised actors that track users only in a given location (e.g., CN but not UK or vice versa). In contrast with UK websites, whose third party providers are mostly US-based, CN is dominated by local third party providers.

RQ3: Do all countries experience the same amount of tracking?

§4.4 also shows UK users are tracked *more* than in China — the dominance of players like Google results in individual players obtaining a large coverage of users' browsing patterns. China's third party ecosystem is more decentralised; which results in diminished visibility and coverage of individual third parties. However, there are *fewer* social third parties targeting UK users than CN users. Also, it is observable that Google manages to obtain non-trivial, albeit diminished, coverage of users in China through some of its domains which are not blocked.

RQ4: Do trackers use traffic discrimination?

I find websites have dynamic strategies to load different trackers over time (§4.3). These strategies are segmented based on the location the user connects from and the type or category of website the user connects to. By combining a more common study of different websites based on Alexa rankings with a study of real browsing histories of a panel of users, it comes to the conclusion that Alexa based studies may be systematically over estimating the amount of tracking that individual users may experience.

4.3 Tracking patterns across countries

This section provides an overview of the magnitude of the tracking problem, and describes patterns common across both China and UK. I study the capability of third-party networks to track individual users over time, and a group of different individuals over a single time period (§4.3.1). Then I study where tracking is more prevalent, by exploring tracking on first party sites in different categories and with different popularity ranks (§4.3.2). Finally, I look at how the time spent on a web site affects the tracking strategies (§4.3.3).

4.3.1 Measuring tracking with overlaps

Trackers derive their power from obtaining a panoramic overview of browsing habits. I extract a measure of this by studying overlaps in the first and third party domains across users and over time. Intuitively, this measure shows the similarity between two sets of browsing behaviours. There are two main applications to this. First, it can be used to measure how much overlap there is between the browsing behaviours of the same user at two points in time [DM14]. Here, this measure can show to what extent a third-party network can track the user across the Web during that period. For instance, if there is a high overlap of third-party cookies across time, the issuer of the cookie will be capable of inferring most of the browsing history of the user during that period. Second, it can be used to measure how much overlap there is between the browsing behaviours of two users (or groups of users). This tells how similar two users are and, when looking at the third-party overlap, it can give a notion

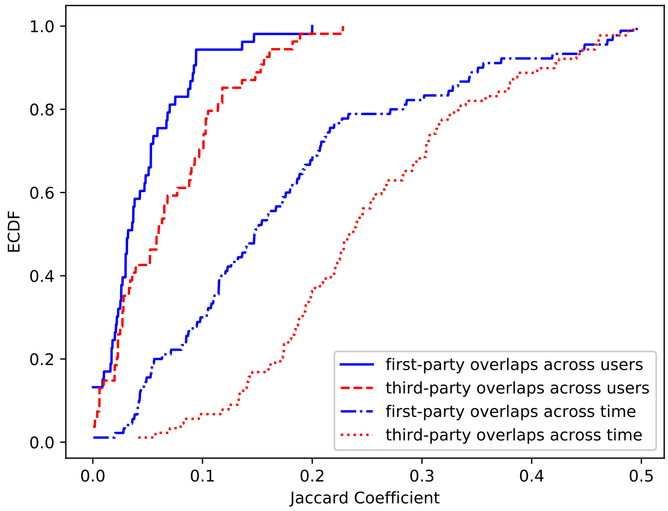


Figure 4.1: Overlaps across users and time in the real-user database. *Third parties obtain broader coverage of browsing habits than first parties.* This is true both for individual users over time and across users within the cohorts I study.

of how well a third-party network can track a population or cohort of users. It can also be used to compare the different user groups by country or location for instance.

In this chapter, Jaccard coefficient is used to measure the overlap/connectivity of third-party trackers between the two separated websites A and B. Formally,

$$Jaccard(A,B) = \frac{\sum overlaps}{\sum websites} = \frac{|A \cap B|}{|A \cup B|} \in [0,1].$$

I empirically observe that there is no correlation between Jaccard coefficient **(not exceed 0.25)** and the number of websites visited by different users in the dataset, implying that the amount of browsing of a user does not create a bias for this measure (at least in the proposed dataset of this thesis).

For each week, I compute the first- and third-party overlaps *across users*. The first-party overlap gives us a notion of how many users land in the same pages. The third-party overlap gives a picture of the extent to which third-party providers learn about similarities among users' browsing histories. I also measure the overlap of first- and third-party domains *across time* for individual users (i.e., the extent to which users revisit the same websites over different weeks, and the extent to which third parties know about the temporal visiting patterns of that user).

Figure 4.1 depicts these overlaps. There are three takeaways: First, there is *higher* overlap in the third-party domain than in the first-party domain, both across different users and for individual users across time. This implies that third-party providers have a more comprehensive overview of browsing habits of individuals and cohorts of users than first party providers. Second, the overlaps of browsing histories over time for individual users is higher than the overlaps of browsing histories across different users, and this holds both for first-party and third-party domains. This implies that third parties are able to track individual users more effectively than tracking cohorts of users, and indicates a degree of specialisation (e.g., due to targeted advertising), whereby a third party may be interested in (or have visibility of) some users but not others. Finally, the difference between the first- and third-party overlap across cohorts of users is not as wide as the difference between first and third parties seen by individual users. This suggests that third parties do not have a massive advantage over first parties in understanding behaviours of cohorts of users. However, it should be noted that some of the largest domains, such as *Google* and *Facebook* act as both first and third parties depending on the context. For example, Google which can be a first party for search queries, is also a third party for analytic (Google Analytics) and for advertising (DoubleClick). As first party, Google has an extensive 64.2% coverage of the browsing histories of the UK user base. Thus some of the most common first parties may have much higher overview of users' browsing histories than third parties.

A different way to estimate the magnitude of tracking is to consider the extent to which trackers are shared among different kinds of websites. To study this, I make use of the categorisation of websites by `Alexa`. Figure 4.2 shows the overlaps in third party domains between websites `Alexa top500` of different categories. The overlaps of third parties among most categories have a Jaccard coefficient in a tight band between 0.2 and 0.4, suggesting that in general, the category of a website does not make a huge difference to the presence or absence of particular trackers. However,

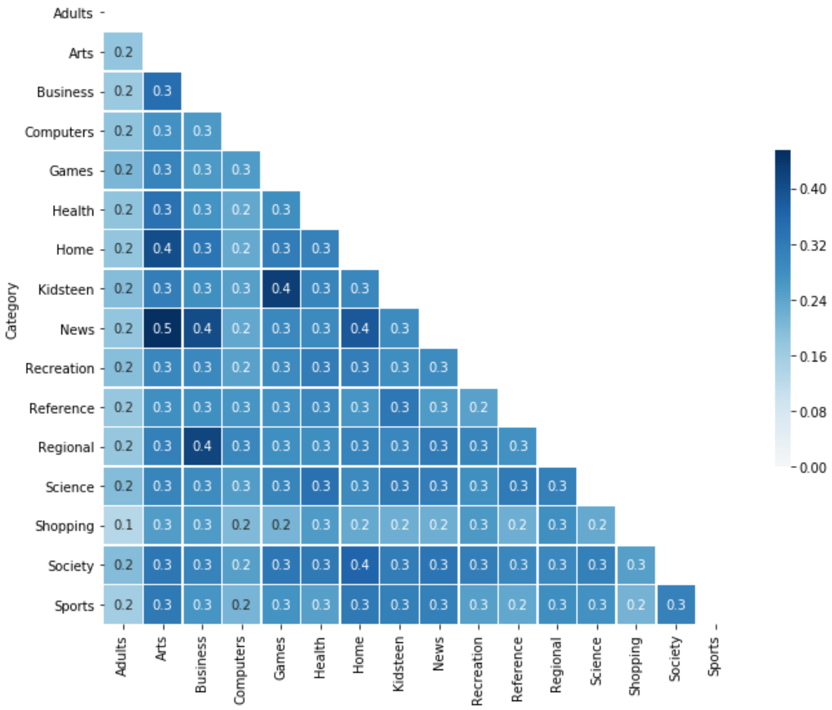


Figure 4.2: Jaccard Coefficient of third parties across Alexa top websites by category.

there are notable exceptions: the highest degree of overlap is between Arts and News, which both largely focus on the digital publishing industry, which has a high Jaccard coefficient overlap of 50%; Also, some pairs of categories with an expected affinity (e.g., Kidsteen (kids & teens) and Games websites, News and Business, or Sports and Games) and have a nearly Jaccard coefficient overlap of 50%. Another notable exception is the category of Adult sites, which have a very different ecosystem of third parties. These sites have a very low (≈ 0.2) overlap with most other categories of websites. This implies a degree of privacy for users visiting Adult websites, as called for by some regulators [AHH16], and maybe a consequence of explicit policies that some large trackers and mainstream advertisers have of not wanting to be associated Adult sites.[3]

Category (#TPs)	Category (#ADNS)	Rank	Category (#Csync)	Rank
News (3156)	News (396)		Sports (207)	↑ 1
Sports (3051)	Sports (392)		Recreation (201)	↑ 4
Business (3057)	Business (336)		Shopping (198)	↑ 8
Arts (2814)	Arts (328)		Business (177)	↓ 1
Home (2763)	Home (300)		KidsTeen (171)	↑ 2
Recreation (2130)	Regional (280)	↑ 3	Home (165)	↓ 1
KidsTeen (2100)	Reference (268)	↑ 7	News (159)	↓ 6
Games (2043)	Society (268)	↑ 2	Arts (144)	↓ 4
Regional (1890)	Recreation (265)	↓ 3	Regional (135)	-
Society (1689)	Science (256)	↑ 3	Games (117)	↓ 2
Shopping (1641)	Shopping (248)		Society (114)	↓ 1
Health (1578)	Games (246)	↓ 4	Computers (111)	↑ 3
Science (1491)	Computers (244)	↑ 2	Health (101)	↓ 1
Reference (1284)	KidsTeen (242)	↓ 7	Science (99)	↓ 1
Computers (1140)	Health (236)	↓ 3	Reference (54)	↓ 1
Adults (1134)	Adults (212)		Adults (12)	-

Table 4.1: Number of third party domains seen in `Alexa top500` per category before (#TP), after **authoritative DNS** (#ADNS) and cookie synchronization (#Csync) disambiguation.

4.3.2 Impact of popularity rank & category

Inspired by the previous result of differences in overlaps between categories of websites, I next characterise to what extent the tracking varies among websites. First, in Table 4.1, I count numbers of third parties in the `Alexa top500` in different categories ranked by order. I establish that News websites have the highest numbers of third-party domains, and Adults the least numbers. In other words, the mainstream and accepted web browsing activity of reading news online has the highest amount of tracking and privacy violation. This confirms previous findings [EN16]. However, a limitation of previous works is that they do not take into account authoritative DNS (ADNS) and cookie synchronisation, where two third parties might open a side channel to share data.

In estimating how much of a user's browsing history a single entity may be privy to, it is insufficient to just use the domain name of the third party — a single entity

[3] E.g., see Taboola's policy https://bit.ly/2Vr9kQ9.

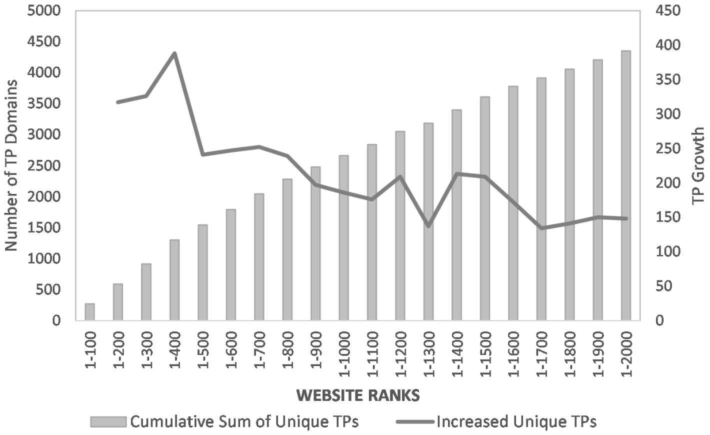


Figure 4.3: Growth rate of the number of third parties and the number of unique third parties in `Alexa top2000` **(a new database extracted from Alexa global top websites)**, divided into bins of 100 sites by popularity ranks.

may simply employ multiple domain names, either to explicitly hide the extent of tracking, or as a result of organic discrepancies arising in domain name usage (such as due to mergers). For instance, Google owns multiple other domain names such as `doubleclick.net` and `google-analytics.com`. To disambiguate such cases, we follow previous work [KW09] and merge third parties if they are controlled by the same Authoritative DNS Server (ADNS).

Secondly, in estimating the loss of privacy,I also take into account possible data sharing through cookie synchronization [PKM19] as a mechanism to establish a "data sharing tunnel" between different third-party vendors. Cookie synchronization can be detected by correlating unique userIDs embedded in cookies stored by different third parties. The methodology in this study follows the guidelines given in [PKM19], reseting the length range of extracting userIDs and reposition the length reduced process to speed the detection. I adopt the twofold restriction on the length of the userID (red dashed line in Appendix Figure 4.4): while decoding cookies and splitting userIDs, which increases the time complexity and saves about 20% of the processing time. Figure 4.4 shows the workflows/model of how I detect the relationship among different third parties.

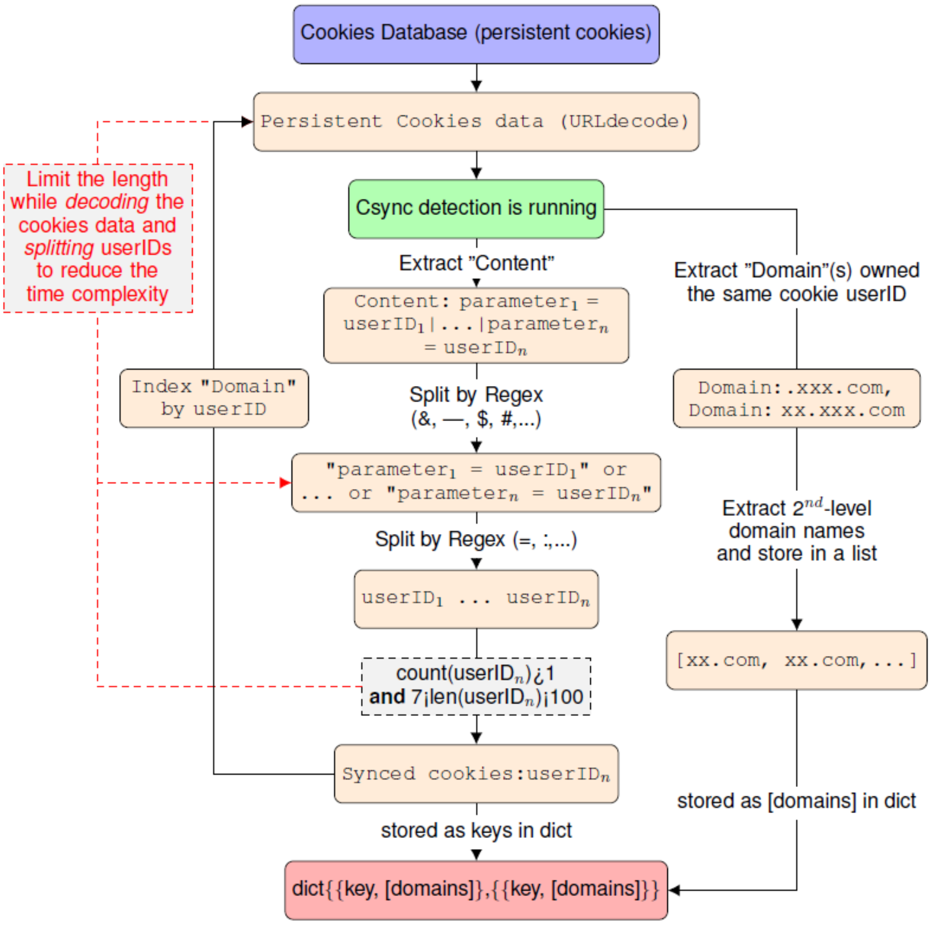

Cookies Database (persistent cookies)
Persistent Cookies data (URLdecode)
Limit the length while *decoding* the cookies data and *splitting* userIDs to reduce the time complexity
Csync detection is running
Extract "Content"
Content: parameter$_1$ = userID$_1$|...|parameter$_n$ = userID$_n$
Extract "Domain"(s) owned the same cookie userID
Index "Domain" by userID
Split by Regex (&, —, $, #,...)
Domain:.xxx.com, Domain: xx.xxx.com
"parameter$_1$ = userID$_1$" or ... or "parameter$_n$ = userID$_n$"
Extract 2nd-level domain names and store in a list
Split by Regex (=, :,...)
userID$_1$ ... userID$_n$
[xx.com, xx.com, ...]
count(userID$_n$)¿1 and 7¡len(userID$_n$)¡100
Synced cookies:userID$_n$
stored as [domains] in dict
stored as keys in dict
dict{{key, [domains]},{{key, [domains]}}

Figure 4.4: Detection of synchronized cookies

Table 4.1 presents the number of third parties after disambiguation together with the relative change in the rank of the category. Assuming that merged entities share data, a decrease in their ranking means that users browsing pages in that category are more prone to be tracked than what was previously reported. I argue that when the diversity of third parties in a set of websites is reduced, single trackers then gain a better overview of a cohort. The data shows more consolidation among trackers in KidsTeen, Health, Games or Recreation after coalescing by ADNS. Note that KidsTeen registers the largest rank decrease. Although websites in KidsTeen look like they have many smaller TP players, these are related entities and each player is bigger than what appears judging other works [EN16]. When looking at cookie synchronisation, I observe that third parties in Sports, Recreation and Shopping categories share the largest number of cookies with other third parties. The content offered in sites under these categories enables targeted advertising and there is a greater incentive to share user's habits through cookie synchronisation. The rest of the chapter presents results after ADNS disambiguation, i.e., only third parties with different ADNS servers are recorded as distinct entities.

Next I ask whether websites of different levels of popularity also have different levels of tracking. Figure 4.3 counts the number of unique third parties added as I go from the most popular websites (`Alexa` ranks 1–100) to less popular ones (up to `Alexa` rank 2000). As I move down the popularity rank, the cumulative number of unique third parties found continues to grow, indicating a vast and well developed ecosystem of third party providers. However, I find that the *number* of new third parties added for each 100 ranks plateaus out after an initial peak caused by a few of the popular websites. As a corollary, this means that the many academic papers which focus solely on `Alexa` ranked websites (e.g., [EN16, Lib15, CKB12, LN18]) may be providing an *upper bound* on the amounts of tracking. For instance, while authors in [EN16] study 1M sites, they only sample top (100) sites in different categories. With some notable differences,[4] the number of third parties is over-approximated when considering sites in different popularity ranks. Equally, real users with browsing habits including specialist or niche-interest websites that are not among the most popular sites automatically tend to have lesser tracking (specially within the advertising industry). I confirm this by looking at the user group, where the number of third parties also plateaus out after an initial peak.

In Figure 4.5, I seek to understand this result further by examining how different categories of trackers are used in websites of different popularity ranks. In most categories, there is not much difference in the relative proportion of trackers of that category among sites of different popularity ranks. However, advertising is one notable exception: the number of advertisers drops sharply after the top 1000 ranks, corresponding to the plateau of Figure 4.5. Thus, the difference observed in the

[4] E.g., Number of third parties remain steady across ranks for Government-related sites; numbers grow for Games and Shopping.

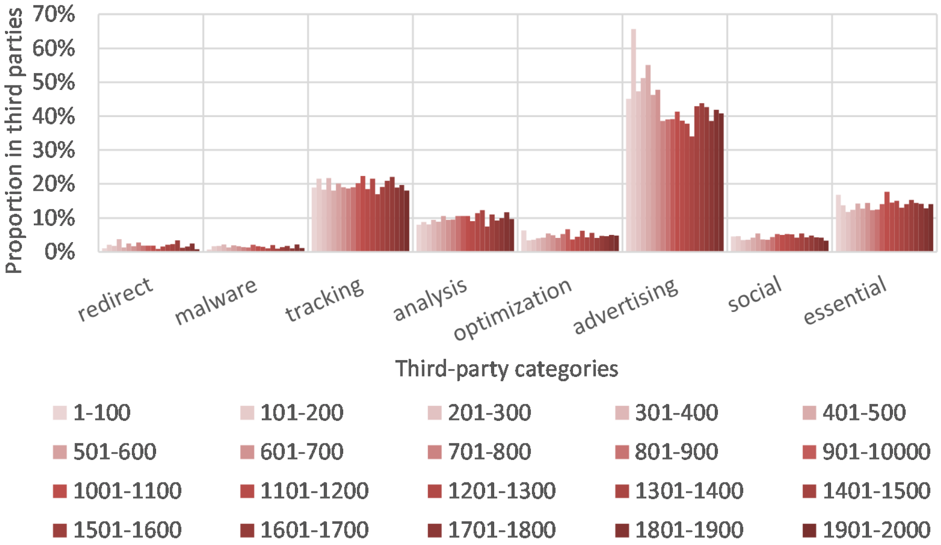


Figure 4.5: Proportion of third-party categories in `Alexa top2000`. **Noted that one domain sometimes supports more than one service, so they do not add up to 1.**

number of trackers may be a result of financial pressures and incentives for online advertising, which pay more for more popular sites, and conversely, are less present in less popular sites.

4.3.3 Impact of loading time

Finally, I simulate an unusual experiment, to understand how tracking may change over time: I load `Alexa top100` websites in 100 Selenium instances and instruct Selenium to record all connections made after loading the site. Each website is to run for seven days continuously and aggregate the observations on the time scale of each minute. This allowed for the computation of the per-minute rate of the increase or decrease in the number of third parties in each third-party category. Table 4.2 shows the daily average increase in the numbers of trackers seen. Overall, I observe mostly positive values, indicating that the number of third parties keeps increasing over time even after several days. The highest rate of increase is seen in the Advertising and Tracking categories of third parties, especially on days 3 and 4. Note that these two categories represent the largest fraction of third-party connections (c.f., Fig. 4.5). This

	redirect	mal	track	analysis	opt	ad	social	essential
day1	0.01%	0.10%	0.21%	0.11%	0.27%	0.33%	0.03%	0.21%
day2	0.02%	0.02%	0.21%	0.09%	0.27%	0.56%	0.04%	0.28%
day3	0.09%	0.09%	0.38%	0.14%	0.25%	0.75%	0.04%	0.23%
day4	0.08%	0.11%	0.69%	0.59%	0.14%	1.11%	0.09%	0.98%
day5	0.04%	0.18%	0.43%	0.45%	0.34%	0.24%	0.09%	0.24%
day6	0.03%	0.07%	0.15%	0.15%	0.31%	0.52%	0.08%	0.17%
day7	0.08%	0.18%	0.22%	0.09%	0.23%	0.43%	0.15%	0.65%

Table 4.2: Alexa Top 100 sites are loaded and kept open for 7 days in 100 Selenium instances. Each entry in the table shows the daily average of the per-minute rate of increase in the numbers of third parties of a particular category. (*mal :malware; opt: optimisation; ad: advertising)

seems to suggest a widespread practice of regular turnover of advertising and tracking third parties. To the best of knowledge, this project is the first to report this behaviour. I investigate how this behaviour varies by country in Sec. 4.4.3.

In a small number of cases, I also observe that the tracker changes over time, over much longer time scales than the 7 day period of the above experiment, but visible in the year-long browser histories of the users. This change in trackers occurs due to a *renaming* of the tracking domain itself, i.e., a change in the domain name of the tracker. To exclude the influence of loading time in the analysis, I open the same websites and record the number of changes in third parties per minute over 30 minutes. I observe that there are changes over time, but these are not very prominent.

This occurs typically as a response to the third party domain being listed on a blocker's list such as uBlockOrigin, a free and open-source[5] browser add-on for the online ad-blocking. In response to this block, I observe that the old domain name is dropped, and a new domain name with a similar sounding name is registered. Figure 4.6 lists a few examples. For example, shortly after `pussl4.com` is included in the uBlockOrigin list, that third party stops getting included in third parties loaded by first party sites, and instead is replaced by `pussl3.com`. Similarly, after `wikia-beacon.com` is blocked by uBlockOrigin, this is moved to `beacon.wikia-services.com`.

[5] Github: https://github.com/uBlockOrigin/uAssets

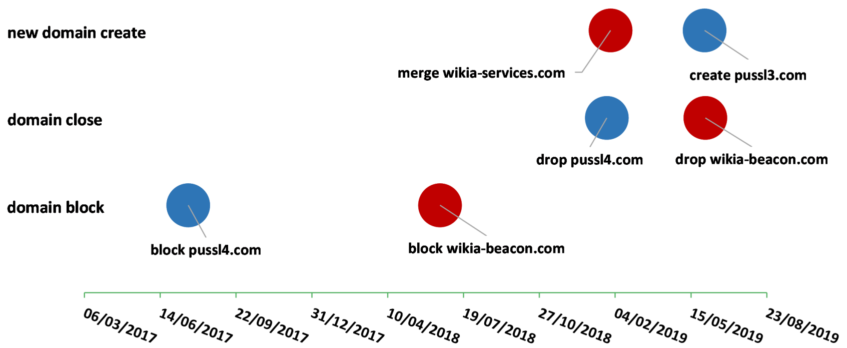


Figure 4.6: Business cycle of third-party domains, with a timeline typically displaying the following pattern: old domain blocked by listing in external database → old domain dropped (stops getting loaded by first parties) → usage of new domain starts (e.g., `pussl4.com` is replaced by `pussl3.com`) or merge with another domain (e.g., `wikia-beacon.com` is merged as a subdomain of `wikia-services.com`).

4.3.4 Key findings

By looking at third parties across different categories of websites, I note that there is a specialisation with higher overlaps between some categories than others. Furthermore, Adult websites have a lower overlap with most other categories, affording a degree of privacy. **According to the collected real-user database, it is shown that** numbers of trackers drop off as the popularity rank of a website decreases — real users, who may visit a significant number of niche interest websites outside the `Alexa` most popular lists tend to see fewer trackers. **The access to only a small number of users makes these niche interest websites difficult to act as an interconnected role in the third-party ecosystem. Therefore, the number of trackers generally decreases as the popularity of the site decreases.** This have resulted on recent works [EN16, Lib15, CKB12, LN18] reporting an *upper bound* on the amounts of tracking.

In response to how third-party networks evolve over time, there is an arms race that shapes the tracking ecosystem. On the one hand, I observe that first-party web sites have complex strategies that evolve over time, including loading different third-party technology over time scales of minutes, hours and days. Some of these strategies are likely motivated by the reactive nature of privacy-aware users that use blockers. On the other hand, I observe that the third-party trackers themselves change over time precisely when their domains are blacklisted by the blockers. This is done to increase survivability as dynamic loading strategies of first-parties will not favour loading third parties that are often blocked by their users.

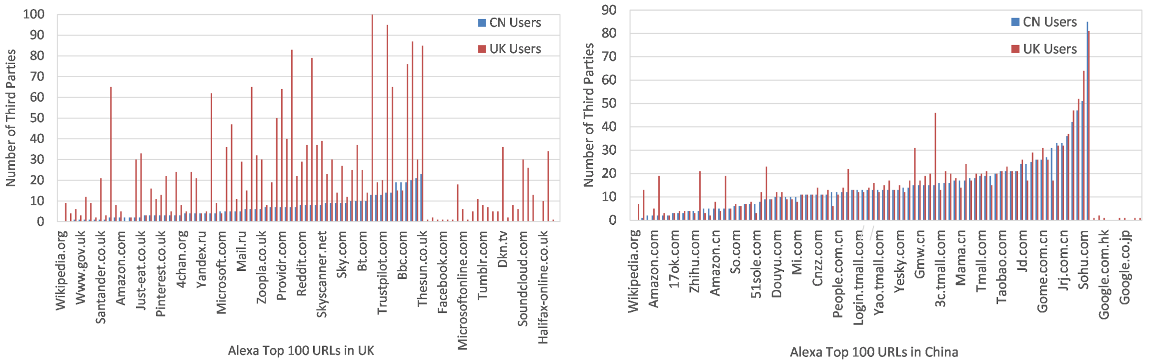
In the rest of the chapter, I focus the attention on the user group to: i) drift away from reporting over approximated (`Alexa`) results, ii) to reduce the impact of popularity ranks and categories, iii) minimise the bias introduced by running experiments with different loading time in a context where real-time bidding might have an affect in the understanding of the tracking ecosystem.

4.4 Country-level Differences

Due to the differences observed between UK and China, in this section I look at third-party technology that operates at a country level. First, I contextualise the study by comparing global third parties with local (CN and UK) third parties in §4.4.1. I then look at the cumulative growth on third parties targeting users throughout one year (§4.4.2). And lastly, I present the top actors across sectors and per country in §4.4.3.

4.4.1 Number of Third Parties in CN & UK

It is common to find websites that own the same second-level domain name in different countries (i.e., under several top-level domain like *.com* or *.uk*). Other sites, like Taobao, one of the highest traffic websites in China, operates two versions of their website homepage in different third-level domains (*'www.taobao.com'* for local users and *'world.taobao.com'* for global users). This is typically used to customise the homepage based on the origin of the users. During the course of the study, I have observed how some of these sites insert different third-party technology on the sites they own. This suggests that some sites might tailor the number and type of third parties based on the location of the user. To verify this, I first study the interplay between the location of the user and the overall number of third parties observed.



(a) UK Top 100 sites visited from UK and China locations.

(b) China Top 100 sites visited from UK and China locations.

Figure 4.7: Number of third parties targeting Chinese (CN) and UK **location** users respectively on Alexa UK (top) and CN (bottom) most popular 100 sites. Top UK sites appear to target UK users more, but top Chinese sites target users from both CN and UK locations equally. (Each figure has 100 sites, but only a selection are labeled to ensure legibility).

Hosting Loc. (CN users)	**x% at loc.**	**Hosting Loc. (UK users)**	**x% at loc.**
China	66.3%	United States	76.6%
United States	24.5%	United Kingdom	7.7%
South Korea	1.8%	Ireland	5.2%

Table 4.3: Top 3 hosting locations of Third Party Domains encountered by real users from CN and UK in the user study. Chinese users are mostly served by localised third parties in China whereas UK users are tracked by US-based third party providers.

I start by connecting from **residential** locations in China (CN) and UK to `Alexa` top 100 websites (first parties) of each country**, respectively**. It is important to note that all sites are loaded in a controlled experiment, where I connect from the CN/UK locations to the same sites simultaneously. I also ensure that a clean browser profile with no previous history of cookies is used when visiting each website.

Figure 4.7(a) shows the number of third parties observed when connecting to `Alexa` **UK** top 100 websites from China (CN) and UK. Observe that users located in the UK see significantly more third-party technology than users in CN. Interestingly, when repeating the experiment with `Alexa` **China** top 100 websites (Fig 4.7(b)), I found that the number of third parties does not vary as much with the location of the user. This suggests that *trackers in UK websites are more location sensitive than in China.*

I explore this systematically in Figure 4.8 by loading each website from the `Alexa` **global** top 2000 websites from locations in UK and CN, and counting the numbers of third parties (TPs) observed. I find that *across websites of different levels of popularity, UK-based users see more trackers than CN-based users.*

This is an unexpected result. I conjecture that this may partly be because users' locations play an important role in advertisers deciding whether to place an ad or not. Other third parties may also have similar reasons. However, an important reason may be that certain third parties are being thwarted by the Great Firewall (GFW) of China, which blocks services such as Facebook, Twitter and Google. None of these domains are seen from the China locations. In summary, *whether because of user demographic characteristics inferred based on locations, or because of GFW, users in China are subject to lesser tracking than users in the UK.*

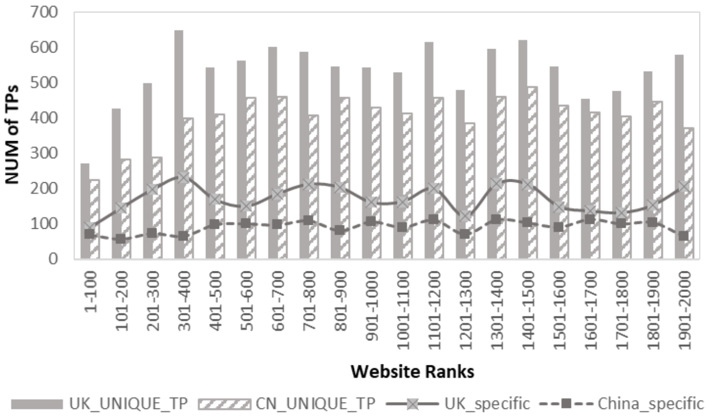


Figure 4.8: Numbers of third parties seen on `Alexa top2000` (global ranking) websites, when accessed from UK and China (CN) locations; and the number of country-specific third parties found only when accessing from one location (UK or CN).

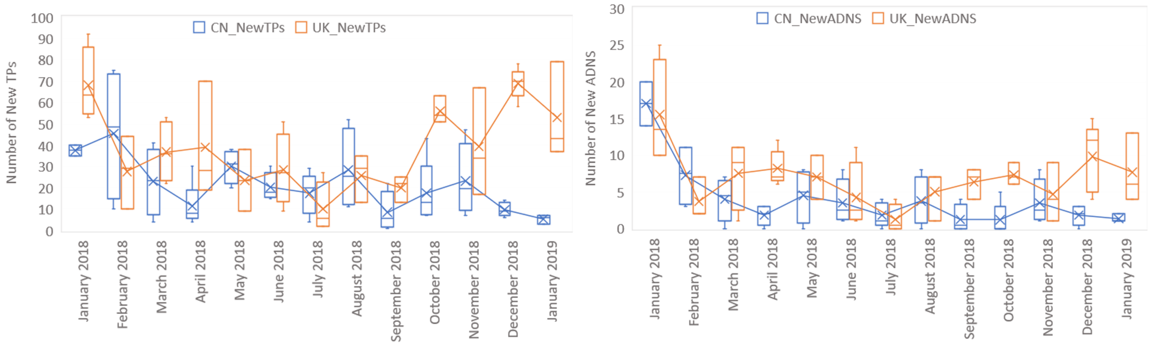
Figure 4.8 also shows that there is a number of third parties which are only seen in the UK and not in CN. In total, approximately 46% of TPs seen in the UK are not seen in CN at all. This should be expected because of the above stated reasons for censorship at the GFW and demographic specialisation. Interestingly, I also observe that there are several TPs which are only seen in the CN and not in the UK. 34% of TPs seen in China are endemic to users in that country. This indicates that *China also has a strong home-grown ecosystem of TPs.*

I explore this further in Table 4.3 by looking at all of the TPs encountered by the user groups in UK and China, and using a `whois` lookup to understand where those TPs are hosted. I find that most (66.3%) of third parties encountered by the Chinese users are located in China, although a significant minority (24.3%) are US-based. In stark contrast, nearly 77% of trackers for UK users are US-based, and only 7.7% are located in the UK. This provides further evidence of China's home-grown third-party ecosystem. The globalised nature of third parties for UK-based users raises important questions about regulations and data management, especially in the wake of GDPR [HS19, ISPL18a].

4.4.2 Evolution over time

I next ask how third parties evolve over time through the lenses of the users. Figure 4.9 shows the cumulative third-party growth over time for UK and CN users during one year. I observe a significant change in the number of new TPs at key times of the year, with spikes in February in China as well as December and April in the UK. These spikes may relate to Chinese New year, Easter, and Christmas respectively, where sites generally make promotional deals from new and different third-party advertisers. Besides, European Union's General Data Protection Regulation (GDPR) was also released during the collection period. Thus, the first data rebound of UK users in Figure 4.9 might be also likely to be affected by new domains that process users' GDPR consents, with Trustarc and OneTrust offering their GDPR consent services starting from March/April 2018 [HS19].

I also do ADNS disambiguation for third-party collection from users, the growth in Figure 4.9(b) is notably lower than in Figure 4.9(a) — judging by the length of the box, I see that the difference between the users becomes smaller. In China, the most apparent ambiguity is in February, while the UK is in December. This means that although the growth of third parties is high during these two periods, most third parties come from the same tracking entity. However, the initial number of third parties in China exceeds those in UK after ADNS disambiguation. This means that the entities tracking Chinese users are *more dispersed*.



(a) Growth rate of new unique third parties.

(b) Growth rate of new unique third parties after ADNS.

Figure 4.9: Growth rate of the number of unique third-party domains/ADNS in the real-user study across one-year tracing.

4.4.3 Localisation and concentration of TPs

Motivated by the findings in the previous section, I look at how the network of third parties is structured globally, as well as in the UK and in China, and how much coverage a single third party can obtain of an individual user's browsing history, or of visits to a well specified set of websites. I first look at a well specified set of web sites – `Alexa top2000` from the global ranking. Then, I use the dataset of real UK and Chinese participants. In both cases, I compute the browsing-history coverage that a single TP provider can obtain. Table 4.4 shows a summary ranked by TP provider for both `Alexa top2000` and real user study.

Case Study 1. Controlled experiment with Alexa top2000

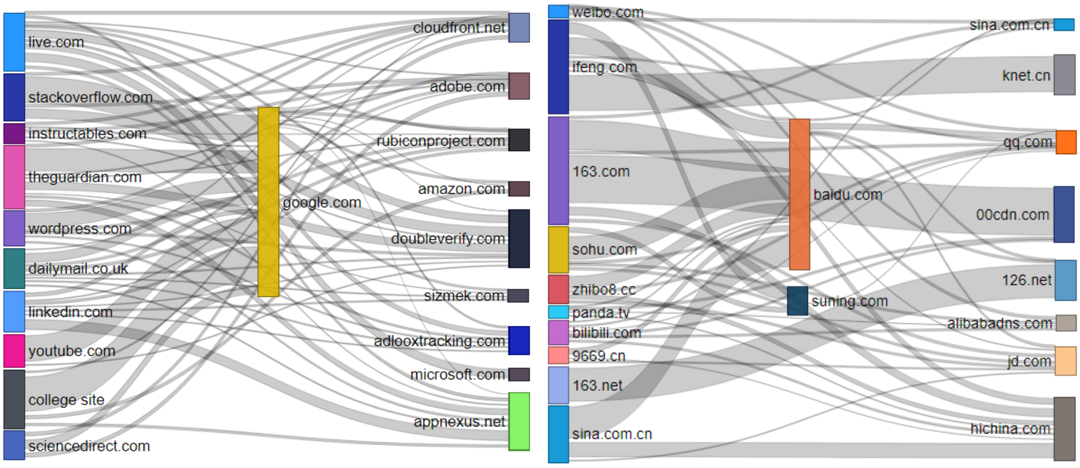
	Proportion of FPs				
Top TPs	**Alexa**	**UK**	**CN**	**AU**	**US**
Google	79.6%	64.2%	11.4%	35.71%	83.85%
Facebook	36.7%	14.9%	-	7.14%	29.53%
Scorecard	20.1%	9.8%	-	-	14.99%
Twitter	13.9%	10.5%	-	2.38%	19.49%
Cloudfront	10.8%	7.2%	1.8%	2.38%	12.11%
Quantserve	7.7%	6.0%	-	-	7.34%
Bing	6.8%	12.5%	-	-	12.68%
Baidu	5.6%	20.9%	46.4%	7.14%	11.06%
Alibaba	-	-	18.2%	11.90%	6.53%
Sina	-	-	13.7%	-	-
QQ (Tenant)	-	-	4.6%	-	-

Table 4.4: Browsing history coverage based on the proportion of first parties (FPs) observed by top third parties (TPs) in **Alexa global top2000**, and in the real-user study from UK, China, Australia and US users.

When looking at the `Alexa` column in Table 4.4, I observe that Google is the provider with largest coverage, and is present in nearly 80% of the `Alexa top2000`. Note again that I have grouped together all known networks owned by Google, like DoubleClick or Google Analytics. The second provider in terms of coverage is Facebook, which is shown to be capable of tracking users out of their site. Facebook has a visibility of 36.7% of `Alexa top2000`. Finally, I can see how other providers like Scorecard, Twitter, or CloudFront have a less dominant share although their presence is still significant. *This indicates a strong concentration of browsing history visibility in the hands of a few top third parties.*

Case Study 2. User study of multi-country participants

I next look at the UK and CN columns in Table 4.4, which represent the third parties seen in the user group. First, I look at third-party networks monitoring UK users**, which have removed CN-UK users**. I observe that Google is able to cover 65% of the browsing history of the users as shown in Table 4.4. This is slightly *lower* than what I observed in the controlled experiment with `Alexa top2000` and corroborates the previous finding (in Sec. 4.3.2) that showed that Alexa-based studies may be



(a) UK users' top10 TPs (#ADNS) traffic flows (b) CN users' top10 TPs (#ADNS) traffic flows

Figure 4.10: Sankey diagram of top 10 third-party websites (after ADNS disambiguation) over traffic of the real-user study (UK on the top, CN on the bottom) topsites. Left bar shows first parties (FPs) and right bars show the third parties (TPs) loaded by the FPs. Each flow represents a loading action, and the width of each flow is proportional to the number of times a TP is loaded by a given FP. (Left bars are top 10 first parties corresponding to the number of loads.)

overestimating the amount of tracking. Note that Baidu is able to observe about 20% of the browsing history of UK users. This corresponds to Chinese users based in UK (referred as CN-UK in §3.2.2).

Next, I look at users in China where Baidu is positioned as the top third-party provider, with a coverage of 46.4%. Interestingly, although Google.cn and DoubleClick ceased operations in China several years ago [LR15], I can see how Google still has access to the browsing history of 11% of the users in China, mainly through other domains owned by Google, which are not blocked. The remaining TP providers are fragmented. This fragmentation might be explained by the low cost of .cn first-year registration domains, which was set to only 1 RMB in 2007 to encourage the development of Chinese websites. This leads to a larger number of TP domains in CN, but with each of them having a smaller overview of the overall market. In the final column (US users), the proportion of TPs in each category shows a high concentration towards Google and Facebook, similar to the percentages found from

Alexa top websites. This may be a result of Alexa rankings being influenced by a large extent by US websites.

Finally, I discuss the provenance of the connections to the most prevalent third parties. Figure 4.10 shows connections from the top 10 first-party websites (on the left side of the Sankey diagram) visited by the UK and CN users to the different third parties after ADNS disambiguation. Note that some Web domains, such as *google.com* and *baidu.com*, can act as either first or third party and they are thus presented in the middle of the Sankey diagram. Here, Google's third-party impact on UK users is *10% higher* than Baidu's third-party impact on the CN users. As a result, Google has more comprehensive user information with higher frequency and reach. Comparing the topology of the UK and CN Sankey diagrams, flows in UK show to be more complex and intertwined. The average first party in UK loads a wider-range of third parties as opposed to China, that is site- or entity-specific. This displays a lower cross-site data leakage in CN over UK and further shreds of evidence that Chinese third parties are site-specific decentralised structures (as discussed in §4.4.2).

Case Study 3. One-year third-party categorisation

I next present an overview of the third-party categories I have seen continuously tracking users for a year (i.e., between 2018 Jan-2019 Jan) in Figure 4.11. I see that in almost all third-party categories, UK providers hold a relatively leading edge. However, in terms of *social* third parties, Chinese providers have a larger network of third parties than UK. Surprisingly, authors of [chi18a] show that "the most widely used social media and content sharing application in the West are banned by the Chinese government", and it can be concluded that China's social media self-marketing has taken up a large space. Note that social media penetration in China mainland is at 71% (Hong Kong is 78%) and UK is only 67% [glo19] of the population. Therefore, frequent activities on Chinese social media attract the interests of relevant third-party providers.

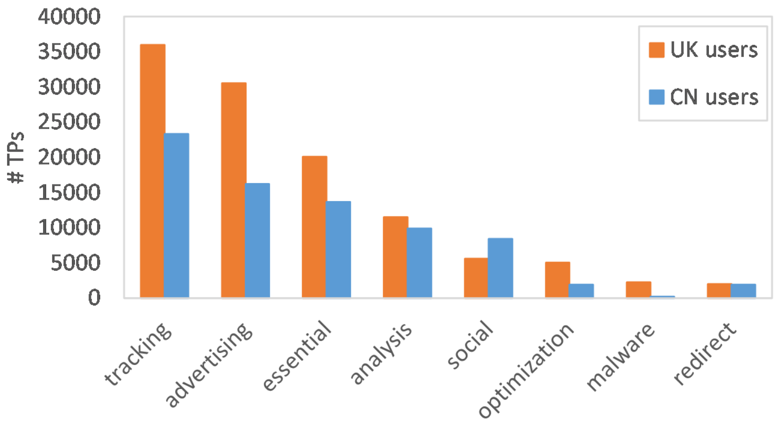


Figure 4.11: Number of third-party categories in actual user research: only *social* third-party vendors display a higher number in China.

4.4.4 Specialisation by category

I also see that, in general, third parties are specialised by sectors. Figure 4.12 shows the type of websites in which the main third party actors have larger coverage. While Google is generally present in all 16 first-party categories, I observe that the presence of other providers changes significantly from one category to another. For instance, Bing (denoted as 'B' in Figure 4.12) is well positioned in the *shopping* category, but it holds poor coverage on websites that are part of the *news*, *games*, or *sports* category. Furthermore, I find that the *adults*, the *reference*, and the *science* category is primarily dominated by Google (labelled 'Gg') alone. In contrast, *arts*, *sports*, and *news* are competitive categories where providers like Facebook ('FB'), Scorecardresearch ('SC') and Twitter ('Tw') stand out in terms of coverage.

4.4.5 Key findings

Although the third-party ecosystem is concentrated in a handful of actors (e.g., Google Facebook and Baidu), I show that there is a degree of specialisation based primarily on: i) the type of sector of the first-party website, and ii) the location of the user. On the one hand, third-party providers that are specialised by sectors and small actors (in terms of overall coverage) can have access to most of the browsing-history

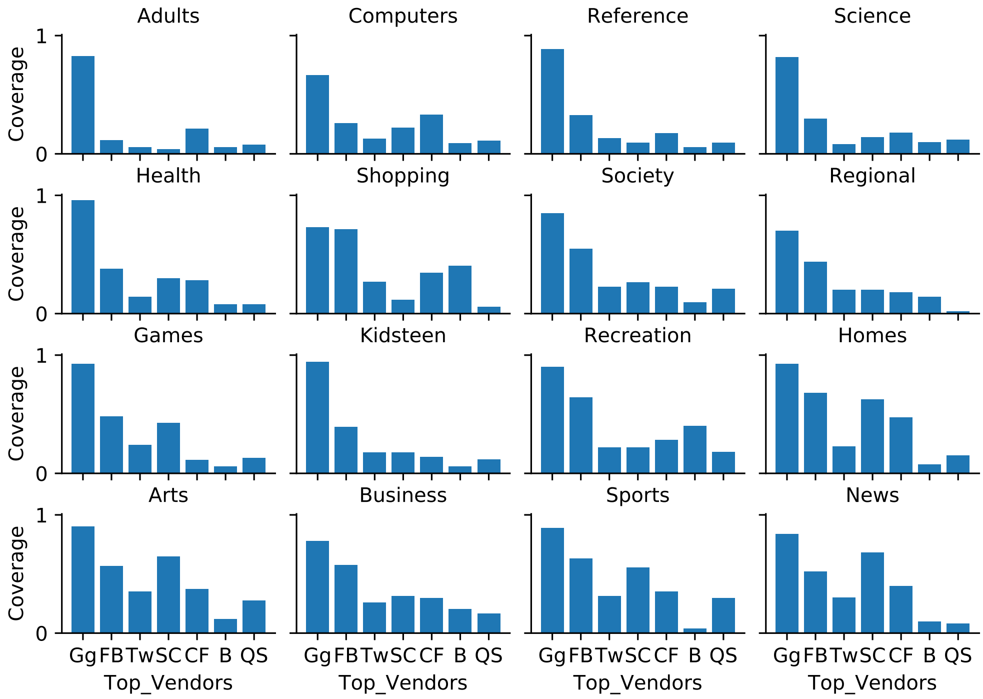


Figure 4.12: Coverage of third party providers among websites in the `Alexa top500` list for each of 16 categories. I only list a selection of the top TPs shown in Table 4.4. (Gg: Google, FB: Facebook, Tw:Twitter, CF: CloudFront, QS: Quantserve, B:Bing, SC: Scorecardresearch.)

of users interested in a given sector (e.g., Scorecardresearch with websites in the *Arts* category). On the other hand, I observe that third-party providers are also specialised by country. However, the vast majority of TP domains in the UK and a smaller portion of TP domains in China are hosted in the US. How user data is processed and where it is located has important regulatory implications. For this reason, large corporations like Google have **placed some of frameworks into practice**[6] to protect cross-country data.

[6] EU-US Privacy Shield Framework: https://bit.ly/2XCgvYn, **but the privacy shield has been declared invalid by the ECJ in July 2020.**

4.5 Discussion

In this chapter, I presented a measurement study that sheds light into the magnitude of the tracking ecosystem in different countries. I built technology to capture the extent to which third-party trackers are profiling users. With a real-user study, I have highlighted the limitations of measurements that rely solely on `Alexa`. I also presented a categorisation of third-party domains that improves the state of the art in terms of performance. All this, together with a set of experiments designed to understand the interplay between the location of the user and the strategies of trackers, showed that this ecosystem is quite complex.

Takeaways. The analysis highlights that first-party web sites exhibit dynamic strategies that change over time, with respect to the location of the user and the *kind* of website the user connects to. In particular, I have observed an important wealth of country-specific trackers as well as trackers that are good at targeting segments of users' browsing histories (e.g., *Shopping*). I have also observed, for the first time, dynamic strategies whereby new third parties continue to be loaded by websites even several days after the initial loading of a website. All takeaways above stem from the `Alexa` dataset. Unexpectedly, I found that UK users see more trackers than China-based users judging by the real-user study, **new third parties are continuously growing faster in the UK even the user visits the same number of websites.** One of the reasons for this is the blocking in China of domains such as Google, Facebook and Twitter, which are not only major first parties but also some of the most important players in the UK third party ecosystem. **The inclusion of UK-China users does not affect the results much, as the site primarily identifies the location visited. It would not contain the all tracker lists from both countries.** Finally, I also observe a relatively larger number of social third parties in China. Finally, being able to study the real browsing habits of a panel of users, I are able to show that studies which rely solely on curated lists of popular websites such as from `Alexa` may be *over estimating* the level of tracking of real users.

Chapter 5

Tangle Factor: Interconnectedness of Third-Party Ecosystem

Summary

As instructed in the previous chapter, when users browse to a so-called "First Party" website (FPs), other third parties are able to place cookies on the users' browsers. Although some of these cookies are placed by FPs, many are placed by third party affiliates (TPs) of the first party sites, for reasons such as advertising, or analysis. Examples include advertising networks such as `adnxs.com`, `amazon-adsystem.com` and `doubleclick.net`, analytical platforms such as `google-analytics.com`, or social media trackers such as Facebook and Twitter. Despite that this practice can enable some important use cases, in practice, these third party cookies also allow trackers to identify that a user has visited two or more first parties which both share the second party. This simple feature been used to bootstrap an extensive tracking ecosystem that can severely compromise user privacy.

This chapter introduces a new metric called the "tangle factor" that measures how a set of first-party websites may be interconnected or tangled with each other based on the common third parties (TPs) share. The insight is that the interconnectedness can be calculated as the chromatic number of a graph where the first party sites are

the nodes, and edges are induced based on shared third parties. On the basis of a network graph constructed with the first/third parties as nodes/edges, disconnect edges to decrease the chromatic number of the graph could be used to achieve the reduction in interconnectivity.

I use this technique to measure the interconnectedness of the browsing patterns of over 2k (2484) users in 102 different countries, through a Chrome browser plugin that I have deployed. The users of the plugin consist of a small carefully selected set of 15 test users in UK and China, and 1000+ in-the-wild users, of whom 124 have shared data with me. I show that different countries have different levels of interconnectedness, for example, China has a lower tangle factor than the UK. This project also shows that when visiting the same sets of websites from China, the tangle factor is smaller, due to the blocking of major operators like Google and Facebook.

Results show that selectively removing the largest trackers is a very effective way of decreasing the interconnectedness of third party websites. I then consider blocking practices employed by privacy-conscious users (such as ad blockers) as well as those enabled by default by Chrome and Firefox, and compare their effectiveness using the tangle factor metric I have defined. The results help quantify for the first time the extent to which one ad blocker is more effective than others, and how Firefox defaults also greatly help decrease third party tracking compared to Chrome.

5.1 Introduction to “Tangle Factor” & Container

There is a growing body of literature that recognises the potential risk of placing the same third party affiliates on different first-party websites. The aggregation of third-party vendors in the user’s browsing history could effectively infer browsing patterns and advertising interests. Then, "targeted" advertisements generated are beneficial to the website owner but at the cost of profiling the visited user.

The existence of Tangle Factor can be clearly seen in a case of that one user simultaneously accesses both www.theguardian.com, and www.thesun.co.uk, since multiple third-party vendors place same set of trackers (e.g., SyncRTB*/DPSync* from Ads.pubmatic.com, usersync from Appnexus, ozone_uid from the Ozone Project) into both of them for correlating the user profile across sites. Therefore, the user-specific information stored in the tracker cookies on different first-party websites is passed to the third party for association, which could be retrieved in accordance with the demand of ad clients.

The case presented supports the perspective that an increase in the number of such trackers could reinforce the entanglement of the weaved browsing network, which is commonly named as the “Tangle Factor” in this chapter. The factor would also positively promote the value of the first-party website. A small scale study by Agarwal, Pushkal, et al. [AJP+20] reached supportive conclusions, finding more sophisticated and intensive tracking techniques leads to a high-priced hyperpartisan websites.

Characterisation of Tangle Factor is important for common online users strengthened understanding of hidden threats within third-party relationships. This work offers fresh insight and visualisation tools to demystify the obscure digital world of cross-site tracking. However, the sophistication of emerging tracking mechanism is always accompanied by the advancement of corresponding countermeasures. As advertisers’ demand for detailed profiles continues to grow, so does the sophistication of third party tracking technologies. Users and many browsers have responded with new technologies to safeguard user privacy. Indeed, this has led to an “arms race”, with users installing ad blockers such as AdBlock Plus[Eye15], uBlock Origin[Hil16] and Ghostery[Cli17], and certain websites (e.g., memeburn.com, englishforum.ch) responding with anti ad-blocking technology[ZHQ+18, MBMC18, MQS17] that refuse to deliver content unless users unblock ad blockers whilst visiting their sites.

The feature in Firefox essentially creates one sub profile for the user in each container, providing distinct sub-databases of cookie storage for that. In practice, a `userContextID` column is added to the cookie database, and only cookie sets matching the context ID of the container are sent to the first-party website by retrieving the `userContextID`. Since each container identity would be kept in isolation, information such as third party cookies are impossible to be shared across containers. Firefox suggests [Moz15] that this can be used to also restrict the third-party sharing and achieve additional privacy, for instance by placing a user’s shopping websites into a separate container from financial websites such as banks and credit cards.

Although the container offer a clean way to create independent cookie storage to accomplish the mission to dissociate first/third parties from each other, it is only effective when the interfering party is manually placed in separate containers. **If the strategy is to place each tab in its own temporary container, it would become resource-wasting and user-unfriendly. The storage of containers might consume plenty of user RAM and make features such as remembering passwords inconvenient. Therefore, the need for automatic allocation of containers exists.** Recently, there is an exemplified add-on undertaken by Firefox to automatically, or by default, provide protection for one of the most prevalent trackers – Facebook. This add-on creates a specialised container [Moz18b] for Facebook. As the name of add-on indicates, it would open Facebook in its own container once installed, and logging out of Facebook for the particular user in other containers, thus automatically preventing Facebook Pixel [Moz18a] and other tracking by Facebook of users who are logged in.

While a solution for Facebook’s tracking is indeed important as it is widely used on many websites, it is not sufficient, as it does not offer protection against other third parties. Furthermore, Facebook is blocked by the Government in countries such as China, and thus is not a major third party in that country [HdTS20]. Thus, solutions are needed that work for other important third parties, as well as for other country and cultural contexts.

Throughout this section, the strategy of separating containers would further cut off the interconnection without affecting the full functionality support of an individual website. To answer the research question that how many containers are required to completely separate different sets of first party websites that existed common third parties with others, I draw on the Tangle Factor of the given set of websites in the browsing history to quantify the interconnectedness of third party cookie ecosystem.

5.2 Methodology

5.2.1 Research Question & Approach

This chapter asks the simple question: if the above **strategy** is applied uniformly across different sets of websites, how many containers will be needed to separate different first party websites that share one or more third parties?

I apply the Tangle Factor metric to three different sets of first party websites. First, I look at the Alexa top-k websites for different values of k, and for two different countries, UK and China. Second, I leverage an ongoing user study[1] which developed a Chrome plugin and collected anonymised browsing histories for a period of one year from a small cohort of users in the same two countries (UK and China). The plugin has since been released in-the-wild, to help users to visualise their own browsing histories. I also provide these users an option to manually send me their histories, and contribute to this study. The third and final set of websites is two months of browsing histories of 2484 users from 102 countries who have decided to voluntarily contribute data to the user study.

I use the tangle factor to understand the third party ecosystem from different vantage points. For example, I visit the top-k most popular websites from UK and China, and find that for the same set of websites (Global top 2K websites according to Alexa), visiting from the UK results in much higher interconnectivity than from China. In other words, the most popular sites have a higher tangle factor from UK locations, i.e., it requires many more separate containers to prevent third parties from tracking browsing of top-k websites in the UK, than in China.

A similar result carries over into actual browsing histories of real users in both countries, which are based on country-specific websites rather than synthetic "top-k" websites: browsing histories of users in China have a lower tangle factor, i.e., are more easily separated, and with fewer containers, than browsing histories in UK. I

[1] This study has been approved by King's College London Research Ethics Committee (**Approval no. MRS-1718-6539**).

then expand to investigate the browsing histories of the in-the-wild users across 25 different countries, and show that the interconnectedness of websites has only a low (-0.0726) correlation with the actual numbers of third parties. Rather, it is the ubiquity of large third party trackers (e.g., Facebook and Google DoubleClick) which are present on a large proportion of websites, that increases the tangle factors.

Consequently, I explore a "what if" scenario where the most common third parties simply did not exist, or were prevented from operating (e.g., through ad blockers). I show that deleting the most common third party trackers, which corresponds to deleting the most common sources of edge creation in the FPIG, is a highly effective approach, and the tangle factor drops very quickly with the removal of the most common trackers. I then use this approach to measure the effectiveness of several ad blockers, and show that uBlock origin is more effective than Adblocker Plus, Ghostery and Ad Guard. I also show that *a)* that the largest containers of non-interfering first parties contain regional and computer-related websites, and *b)* using Adblocker Plus and uBlock origin results in UK interconnectivity dropping to levels similar to that of China. This lower interconnectivity is quantified in terms of the increase in the sizes of the largest set of non-interfering first parties (i.e., the size of the largest container when separating different first parties).

5.2.2 First Party Interconnectivity Graphs

In this section, I mainly take the data provided by the Thunderbeam plugin, and identify different third parties after some subtle disambiguation. After achieving third parties using the techniques mentioned in §3.2.1, I then consider all the third parties of a given set of first parties. I then model this as a graph, where the nodes are the first party websites, and edges are drawn between two nodes if the corresponding first party websites share a third party (after third parties are merged using the disambiguation discussed above).

5.2.3 Calculating Tangle Factor

The tangle factor of a given set of websites, i.e., a given FPIG, is calculated by computing an assignment of FPs to different containers, whilst respecting the restriction that two FPs that share a TP (i.e., two nodes connected by an edge in the FPIG) must be placed in different containers.

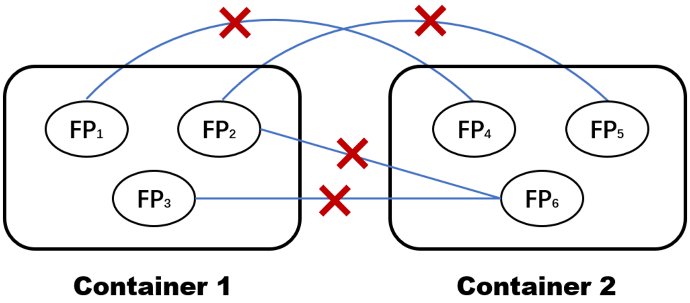


Figure 5.1: Restrictions in this first-party (FP) model example **(The red cross between each line refers to the cutoff of interconnections between containers)**:
① FP_1 cannot be in the same container with FP_3;
② FP_2 cannot place with FP_4 and FP_6;
③ FP_6 cannot be with FP_5.

Effectively, this corresponds to a vertex graph colouring problem, where nodes with the same colour can be placed within the same container, and nodes which share an edge must be assigned different colours. Figure. 5.1 illustrates how a particular set of edges between different first parties would give rise to a container assignment. The minimum number of colours for the vertex colouring problem, i.e., the minimum number of containers needed in the FPIG, is the tangle factor of a given FPIG.

5.2.4 FPIG datasets

I apply the above methodology to obtain tangle factors for several different sets of first parties. First, I use an automated browsing system, built on top of Selenium [EN16] running in non-headless mode, and collect the cookies set by the Alexa[2]

[2] https://alexa.com

User Group	1^{st} **party sites**	3^{rd} **party cookies**
UK Users	8416	113,003
CN Users	6144	74,313
Total	14,827	187,316

Table 5.1: Year-long (Jan 2018 to Jan 2019) data collection from 15 users in UK and China

top-k most popular websites. Specifically, I programmatically visit the country-specific `Alexa top500` sites in UK and China, and also the global `top2k` websites from UK and China locations, in order to obtain a comparison between the number of containers required in both countries. The degree of demand for containers is taken as an indication of the degree of third-party risks in different countries. Further, I visit the UK `top500` websites after installing different ad blockers (uBlock Origin, Adblock Plus, Ghostery and Adguard Adblocker) on two different browsers (Chrome and Firefox), to understand the privacy protection provided by different ad blockers.

To complement this dataset, data are collected from real users who consented to support this work by providing anonymised browser histories[3]. The users come in two cohorts. First, I have a small cohort of 9 users in the UK and 6 users in China, whose browsing activities were collected for a year-long period from Jan 2018–Jan 2019. Altogether, these users have visited around 15k first-party websites across one year, involving over 187k third-party domains (Table 5.1). Since the release of the official version of the add-on in Chrome Web Store as "Thunderbeam-Lightbeam for Chrome", there has seen over 1000+ installs of the plugin by Feb. 7^{th}, 2020. This plugin offers users the option of submitting their data to the study. Through this mechanism, I have collected two-months of data from 2484 users, who collectively provide me a picture of the third party ecosystem from 102 countries, as detailed in Table 3.3 and 3.4.

[3] This research is conducted under the university's Ethics Approval no. MRS-1718-6539. and the data protection review by Chrome Web Store

5.3 Container demand by country

For any two sets of websites, the set with the lower tangle factor (i.e., the number of separate containers needed to ensure that a common third party does not learn about two first parties visited) is the one with the better privacy, and with less powerful third party tracking practices.

I begin by exploring the tangle factor of popular sets of websites from different countries. I then go on to show that tangle factor depends to a large extent on the country rather than actual numbers of third parties involved.

5.3.1 Tangle factors of `Alexa` topsites visited in China and UK

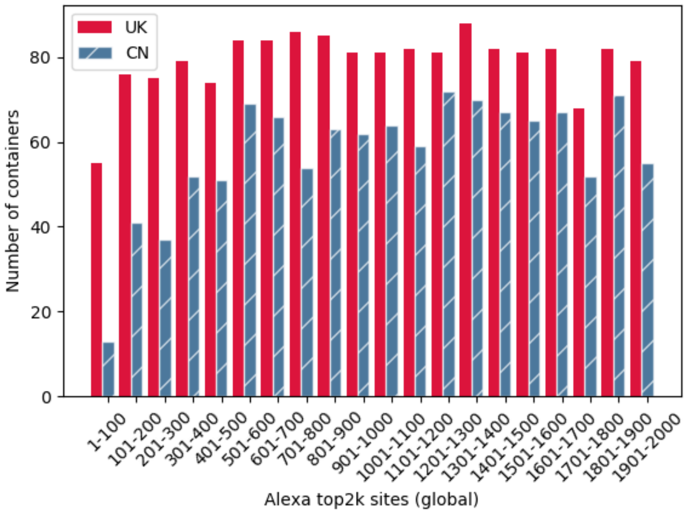


Figure 5.2: Tangle factors for `Alexa` top2k global websites visited from UK and China vantage points.

Since worldwide charts typically use the top 100 as the leading echelon, and Alexa's top 100 sites focus on separate, more specific services and generate a lot of valuable content on a regular basis, this project picks up each 100 sites into a bin. Therefore, this section begins by binning the Alexa top 2K websites (global ranks) into sets of 100 websites (i.e., the first bin has websites ranked 1–100, the second bin has websites ranked 101–200, and so on).

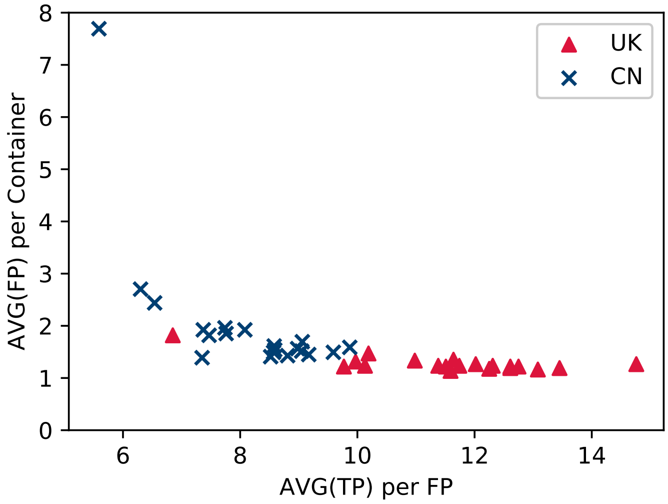
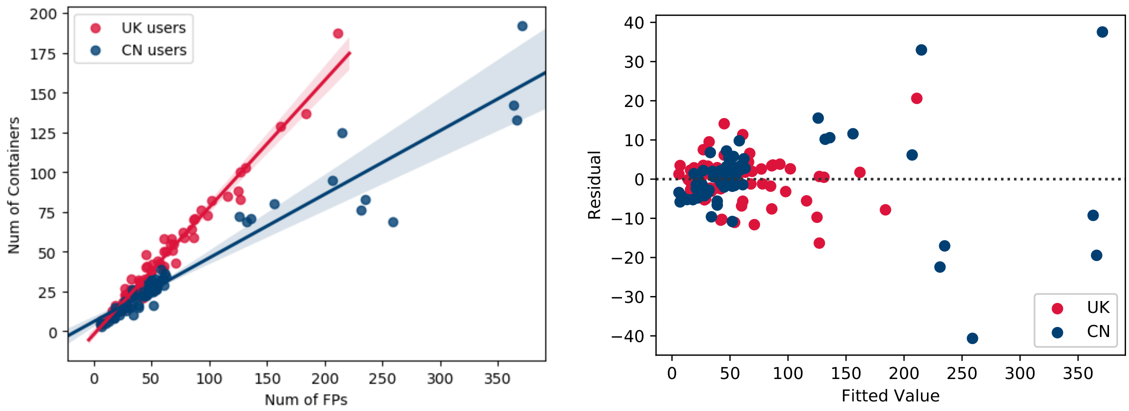


Figure 5.3: Linear correlation between TPs/FP and FPs/Container. Each point represents the average TPs/FP and FPs/container for one of the different ranked bins depicted in Figure 5.2. The Pearson correlation coefficient between TPs/FP and FPs/Container in China is -0.681 and UK is -0.728

I then ask whether users from different countries experience different levels of tracking, by accessing these websites from vantage points in the UK and China, using the automated browsing system (cf.§5.2.4). Figure 5.2 shows the tangle factors of these sites for different rank bins. I find that overall, even when visiting the same sets of websites, users visiting from China are "less tangled" than users visiting from UK, i.e., need fewer containers. This is especially true for the most popular websites (ranks < 500). This is possibly because the Great Firewall of China blocks websites such as Facebook and Google, which are among of the most prevalent of trackers of Western countries [HdTS20]. The scatter plot of Figure 5.3 shows that indeed, the numbers of third parties per website is lower when visiting the Alexa global top2k websites from China than from the UK (blue points are largely to the left of red points with one exception), and that there is a strong (anti-)correlation between number of third parties loaded, and the number of containers needed for the trends among both countries.



(a) Regression between container numbers and the weekly FPs by each user with 95% confidence interval. ($r^2_{UK} = 0.9688$; $r^2_{CN} = 0.9162$)

(b) Residuals

Figure 5.4: Relation between numbers of first parties visited, and tangle factors (numbers of separate containers required) for weekly browsing histories of users from UK and China.

5.3.2 Container shareability in 102 countries

The lower tangle factor in China means that Chinese users need fewer containers to keep their browsing habits private from third parties. Figure 5.5 confirms this, showing that across the duration of the entire year of the study, Chinese users can pack more first parties into each container, without compromising privacy.

I expand on the above observation, and turn to the in-the-wild users of the Thunderbeam plugin (cf. Table 3.3 and 3.4). I ask how many first parties can be packed into containers (on average) for users in different countries around the world.

Figure 5.6 shows the results. In some countries such as China and Singapore (Figure 5.6(a)), it is possible to pack many more first parties into each container, whereas in others, the first parties tend to have common third parties, and therefore need to be placed in separate containers. I simultaneously plot the average numbers of third parties used by each first party, and the figures show visually that there is not necessarily a correlation. Note that this is also the case even when I discount countries with small numbers of users (countries with fewer than five users are greyed out). Figure 5.6(b) confirms the above visual observation with a scatter plot and corresponding Pearson

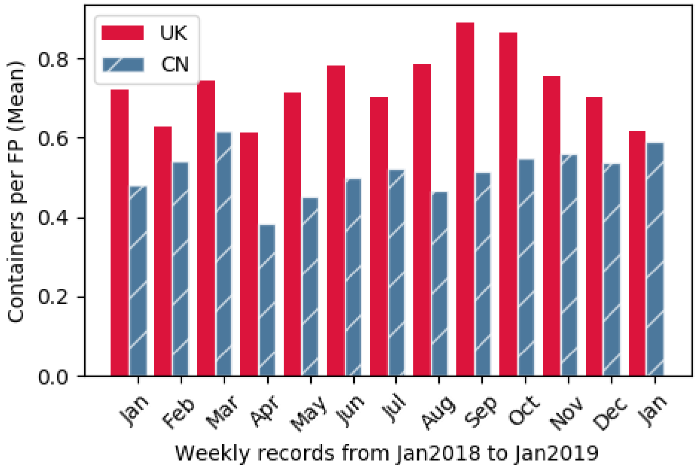


Figure 5.5: Average number of FPs stored in each container from Jan. 2018 to Jan. 2019 after ADNS disambiguation, based on weekly browsing records of UK and China participants.

correlation calculation, that shows there is no strong correlations between higher numbers of third parties used and the numbers of first parties that can be packed into a container. Rather, it is the *interconnectedness* of the first parties, i.e., the numbers of third parties shared, that determines whether or not two first parties can be placed into the same container.

5.4 Restriction of Tangle Factor

The results of the previous section lead me to consider ways to decrease the interconnectedness of the third party ecosystem in a given setting. I consider three options: First, I consider how effective are the default "content blocking" protections of different browsers, comparing Firefox and Chrome (§5.4.1). Then I consider what happens if browsers like Firefox generalised from the current special purpose "Facebook" containers, and instead just removed or blocked the top most prevalent third parties (§5.4.2). After this, I consider user interventions, such as installing ad blockers, which removes certain third parties. Each ad blocker removes different third parties based on their lists, and I use the tangle factor as a metric to understand which ones are the most effective (§5.4.3). Finally, in §5.4.4 I consider how well different categories of websites are able to co-exist with each other in the same container, without information leakage.

5.4.1 Chrome vs. Firefox

To begin with, I assess the default levels of privacy protection provided by two popular browsers - Firefox and Chrome. In particular, many casual or non privacy-conscious users may not install and deploy extensions such as ad blockers, which I consider later. However, even without installing extensions, there is a possibility of using mechanisms such as "Do Not Track" in most modern browsers, including Chrome[4] and Firefox[5], although it is known that Do Not Track is not very effective in practice [Hof19, Bor13]. Compared with Chrome, Firefox provides users with a collection of additional privacy protection features known as content blocking[6]. Users could turn on strict content blocking in Firefox even without installing extensions, and prevent more harmful practices.

To test the effectiveness of these practices, I visit the Alexa `top500` websites in an automated fashion, using Chrome and Firefox with the default settings. I then check the tangle factors, i.e., the number of containers required to separate third parties after the browsers have done their blocking (in the case of Firefox with content blocking), and taking into account the decrease in tracking due to the "Do Not Track" option.

Table 5.2 shows the results. As expected, "Do Not Track" barely has any effect at all. It is merely a request to websites not to track, and if a site chooses not to respect it, there is no effect whatsoever on tracking. However, "content blocking" in Firefox decreases the number of containers required by about 17%, from 410 to 339 containers, because it executes additional blocking on the browser side. However, neither of these options

[4] Turn "Do Not Track" on or off in Chrome: https://support.google.com/chrome/answer/2790761?co=GENIE.Platform%3DDesktop&hl=en

[5] https://support.mozilla.org/en-US/kb/how-do-i-turn-do-not-track-feature

[6] Content blocking: https://support.mozilla.org/en-US/kb/content-blocking

Num (Containers)	**Chrome vresion 80**	**Firefox version 70**
Original	408	410
Do Not Track	405	409
Strict Content Block	×	339

Table 5.2: Number of containers required on Chrome and Firefox, when employing no extensions, but enabling privacy controls available by default.

is as effective as employing an ad blocker such as uBlock origin (which, as I will show in §5.4.3, results in a 40% reduction in number of containers needed).

5.4.2 Effectiveness of removing top third parties

Next step up in terms of user effort, Firefox has an interesting "suggested" add on that isolates Facebook logins from other websites. As Facebook is one of the most commonly used trackers, this is highly effective in decreasing overall levels of tracking. In this subsection, I ask, as a "what-if" scenario, what would happen if all the top-k trackers were removed or blocked by default.

This is visualised in Figure 5.7, which shows the First Party Interconnection Graphs (FPIG) of the Alexa Global top 500 websites, when these websites are visited from UK and China respectively (similar to the setup of Figure. 5.2). The graphs are drawn using a force-directed layout algorithm, with lays out the most connected nodes as the central core, and largely isolated nodes as a ring around the edges. The figure shows how even removing only the top 20 third parties can drastically decrease the tangledness of the FPIGs.

This informal but visually clear result is formalised in Figure 5.8, which shows how the tangle factor progressively decreases when visiting the global Alexa top 500 websites from UK and China, but with the top third parties removed. Initially, UK users need nearly 408 containers for the 500 websites, whereas CN users need 227. However, after removing just the top 50 third parties, the number of containers required drops to 9 and 8 respectively. Thus, *the third party ecosystem is highly interconnected mainly* *because of the predominance of a few large third parties. Protection against these large players can greatly decrease the extent of third party tracking in today's web ecosystem.*

5.4.3 Assessment of ad blockers

Next, I use the same technique as above to determine the effectiveness of ad blockers. I visit the Alexa top 500 websites from a UK location, using Selenium in non-headless mode. I perform this experiment by installing four different popular ad blockers in turn: uBlock Origin, Adblock Plus, Ghostery and Adguard Adblocker. The selection of these four popular ad blockers is based on the recommendations in [AZXW19].

In Figure 5.9, I explore the performance of these ad blockers and show the percentage decrease in numbers of third parties as well as the tangle factor or numbers of containers required, when those ad blockers are deployed. This shows that uBlock origin performs the best, with nearly 60% reduction in raw numbers of third parties, and over 40% reduction in the number of containers required.

Both uBlock Origin and Adblocker Plus use Easy list for the list of third parties to filter. However, different default privacy settings could lead to different degrees of protection. In addition to Easy List, uBlock applies additional filters from Easy Privacy, Malware domain list and Peter Lowe's tracker list[MGF19]. These are enabled by default; thus, even if the user installs uBlock Origin without any custom settings, protection is provided by default. In contrast, Adblocker Plus takes a more moderate approach, and also allows some acceptable ads [AdG19] (e.g., ads that comply with "Do Not Track" or those generated from the same origin as the first-party site), which results in the slight increase in the number of required containers. Adguard Adblocker's poor performance in stopping third parties requests seems to be related to the fact that it hides ad elements after loading the entire site, rather than pre-blocking ad elements[Adb18].

5.4.4 Interconnectedness across web categories

Finally, I ask how different categories of websites are able to co-exist with each other, given the above kinds of interventions. I use Alexa's categorisation of the top 500 websites into 16 global categories, and examine the distribution of these categories in the *largest* of the containers that may be formed after separating websites that share common third parties.

Fig. 5.10 shows the results. The first two columns show the category distribution of the contents of the largest container, when first parties from the Alexa top websites of China and UK are separated into containers based on shared third parties. The remaining four columns show the category distribution of the largest container for the the UK top 500 websites, when different ad blockers are applied in order from the least effective (Adguard) to the most effective (uBlock origin)[7].

The CN500 column has a larger number of sites (94) than UK 500 (61), as tracking is less evolved in China. However, as more intrusive ad block extensions are introduced, the size of the largest container increases even for the UK500, and with both Ad block Plus (ABP) and uBlock origin, the largest container for UK is comparable to or larger than the largest container for China.

Looking across the categories, I find that the largest proportion of sites are those related to computers in both China and UK. Regional websites in the UK also track less and are therefore more easily incorporated into this large container.

5.5 Discussion

In this chapter, I considered the interconnectedness of the third party ecosystem, using vantage points from the UK and China to visit Alexa top 500 websites, and relying on real browsing histories two cohorts of users: one, a carefully selected panel of 15 users, 9 from the UK and 6 from China, the other a set of 124 users from 25

[7] I do not apply Ad blockers to the CN 500 websites, as the ad blocker lists are not adapted for Chinese websites.

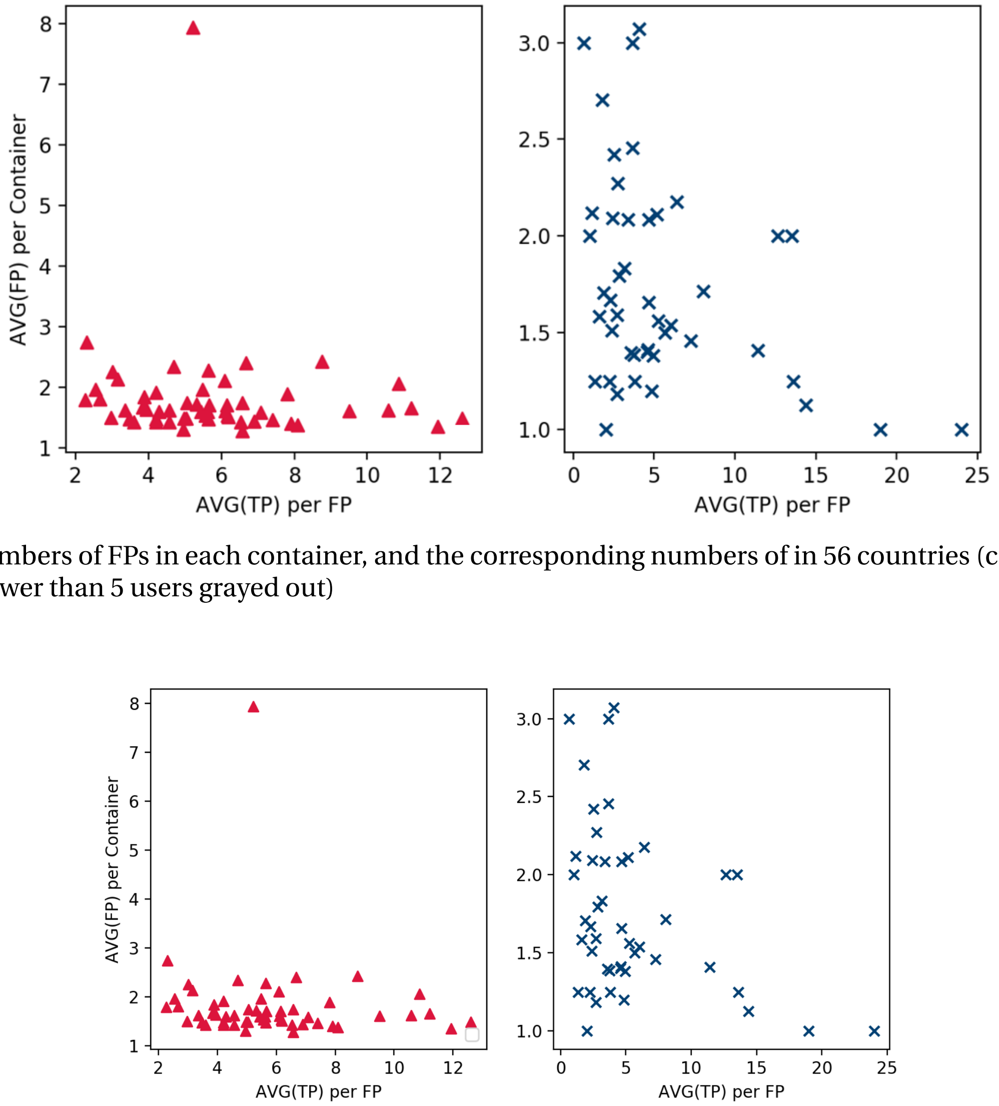
different countries in the world who have chosen to donate data to the research project. I believe I am one of the first to use data from the browsing histories of real users. Note that I do not collect demographic information about users due to privacy considerations.

I introduced a novel metric, which called the “tangle factor”, to measure the interconnectedness of the third party ecosystem. The tangle factor is based on the insight that if I were to create a “first party interconnection graph” by drawing edges between first party websites which share a third party, the websites on either side of edge would share one or more third parties and therefore need to be placed in separate containers to prevent tracking. Thus, the minimum number of colours needed to colour this graph, i.e., its vertex chromatic number, also represents the number of containers needed.

The chromatic number of the graph could be calculated by the greedy algorithm. The first vertex needs to be colored with the first given color. Then, the remaining vertices (V -1) would be one by one considered. If the same color is not used to color any of the next picked vertex’s adjacent vertices, it would be colored with the lowest number color. If one of adjacent vertices are using the same color, then we will select the next least numbered color. If all the previous colors have been already used, then a new color will be used to fill or assign to the currently picked vertex. The greedy algorithm gives an upper bound, and the optimal solution could be determined by computing each vertex.

Using this metric, I showed that when visiting the same websites, users from Chinese locations are less tracked than users from UK locations, likely due to automatic blocking of major trackers like Google and Facebook from the Great Firewall of China. I also showed that this result carries over into the actual browsing histories of the panel of users, which are based on country-specific websites rather than global most popular sites considered in the synthetic evaluation.

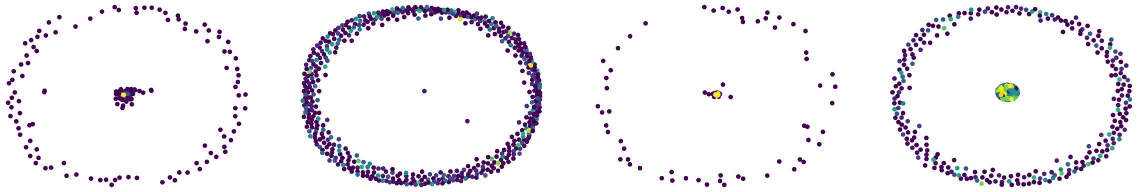
I then used the tangle factor metric to assess the effectiveness of different methods of blocking trackers. I showed that blocking the top 20 trackers alone is sufficient to bring down the interconnectivity greatly, and only 9 containers, instead of 400 containers, are needed in the UK to separate out Alexa top 500 first parties that share a third party. I also used the tangle factor metric to compare ad blockers and showed that uBlock origin works better than others such as Ghostery, Ad blocker plus and ad guard. I also showed that the default protection offered by Firefox is better than that offered by Google Chrome. These measurements are intended as proof-of-concept and the method can be expanded to compare other content blocking and protections apart from the ones I consider in this chapter. These results provide quantitative evidence that the third party ecosystem is highly interconnected mainly because of a few large players, and protection against these can greatly decrease the extent and impact of tracking on the web.



(a) Numbers of FPs in each container, and the corresponding numbers of in 56 countries (countries with fewer than 5 users grayed out)

(b) Scatter plot of number of TPs per FP and number of FPs per container; in red for users in all 56 countries, and in blue for users in 45 countries with more than 5 users

Figure 5.6: Comparison between the average number of TPs per FP and container per FP in countries.



(a) Initial layout of `top500` (CN)

(b) Layout after removal (CN)

(c) Initial layout of `top500` (UK)

(d) Layout after removal (UK)

Figure 5.7: Force-directed layout of the FPIG of the Alexa Global Top 500 websites visited from a Chinese Location ((a) and (b)), and visited from the UK ((c) and (d)). The inner core is highly connected, and the outer ring is largely isolated nodes which can share a container. Nodes which can share a container are given the same colour. (b) and (d) show how the initial layouts in (a) and (c) for CN and UK respectively improve with many more isolated nodes after the top 20 third parties are removed.

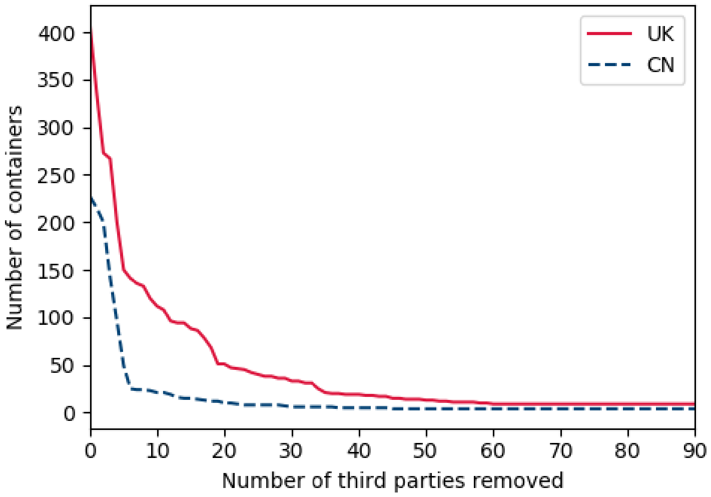


Figure 5.8: Decrease in tangle factor as the top trackers are removed or blocked. After removing about 50 top third parties, UK and China respectively require only 9 and 8 containers, as opposed to initial numbers of 408 and 227 containers.

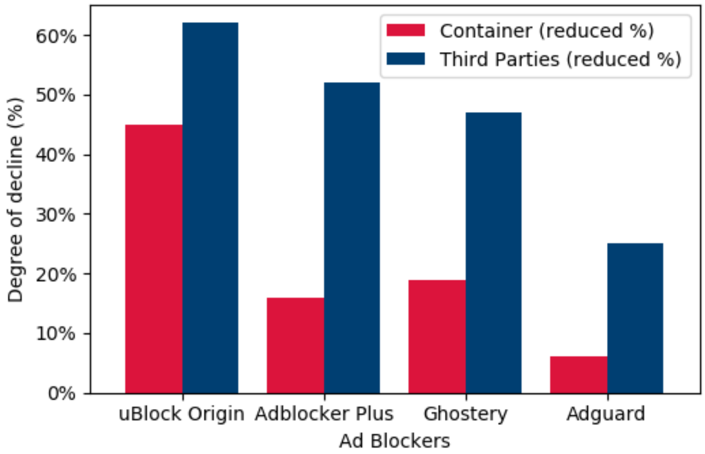
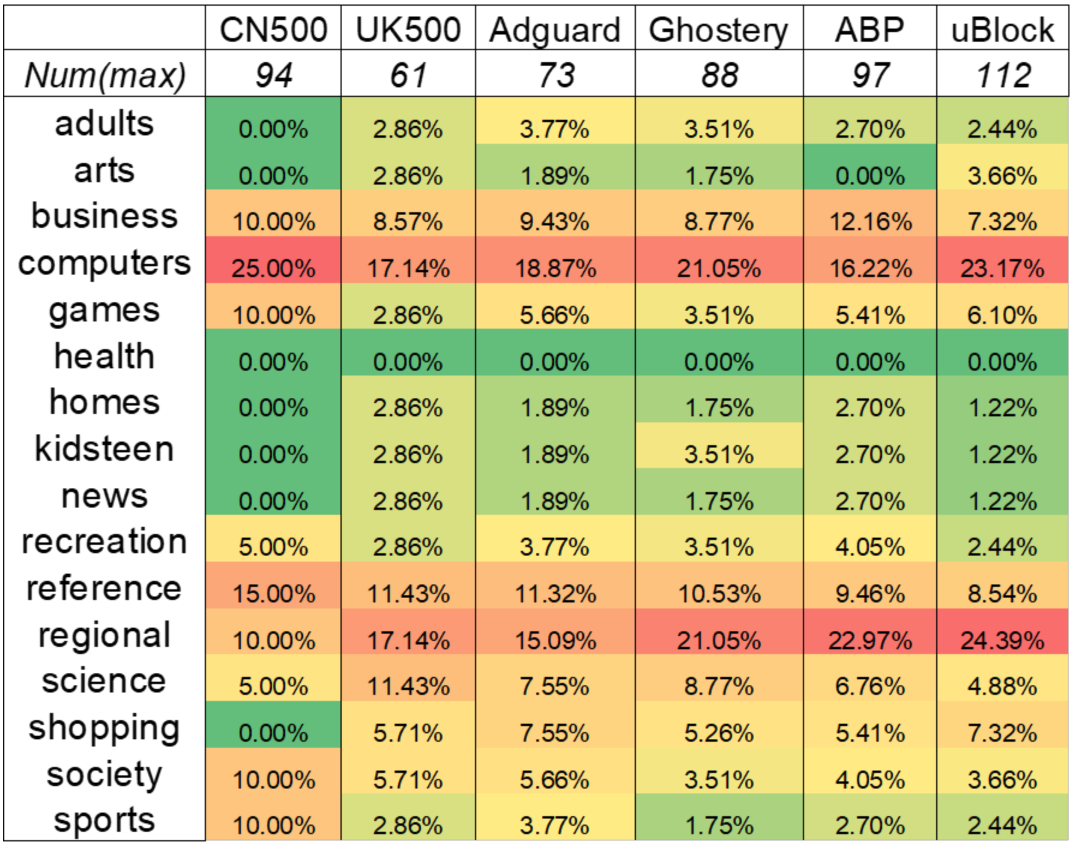


Figure 5.9: Percentage of decline in the number of containers and third parties, when users apply different ad blockers to visit Alexa `top500` websites from a UK location. (Larger decline is better).



	CN500	UK500	Adguard	Ghostery	ABP	uBlock
Num(max)	*94*	*61*	*73*	*88*	*97*	*112*
adults	0.00%	2.86%	3.77%	3.51%	2.70%	2.44%
arts	0.00%	2.86%	1.89%	1.75%	0.00%	3.66%
business	10.00%	8.57%	9.43%	8.77%	12.16%	7.32%
computers	25.00%	17.14%	18.87%	21.05%	16.22%	23.17%
games	10.00%	2.86%	5.66%	3.51%	5.41%	6.10%
health	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
homes	0.00%	2.86%	1.89%	1.75%	2.70%	1.22%
kidsteen	0.00%	2.86%	1.89%	3.51%	2.70%	1.22%
news	0.00%	2.86%	1.89%	1.75%	2.70%	1.22%
recreation	5.00%	2.86%	3.77%	3.51%	4.05%	2.44%
reference	15.00%	11.43%	11.32%	10.53%	9.46%	8.54%
regional	10.00%	17.14%	15.09%	21.05%	22.97%	24.39%
science	5.00%	11.43%	7.55%	8.77%	6.76%	4.88%
shopping	0.00%	5.71%	7.55%	5.26%	5.41%	7.32%
society	10.00%	5.71%	5.66%	3.51%	4.05%	3.66%
sports	10.00%	2.86%	3.77%	1.75%	2.70%	2.44%

Figure 5.10: Proportion of the first-party websites from different Alexa web categories in the largest container

Chapter 6

Cookie Classification

Summary

In the wake of the 2011 EU cookie directive, the UK International Chamber of Commerce (ICC) suggested a four-part categorisation of cookies—strictly necessary, performance, functionality and targeting/advertising cookies. This categorisation is now widely used on the web, especially on websites using the English language. Thus, for effective implementation of policies like the EU cookies directive its necessary to divide browser cookies into one of these categories. Open-sourced platforms like Cookiepedia use data from the tracking industry to divide millions of cookies into these very categories for further usage. However, it is not clear if the cookies categorised by Cookiepedia indeed cover the sites that users normally visit today. In this chapter, I aim to answer this question.

I start with over 37 million cookies collected in Cookiepedia, which is currently the most comprehensive database of cookies on the Web. Cookiepedia provides a useful four-part categorisation of cookies into strictly necessary, performance, functionality and targeting/advertising cookies, as suggested by the UK International Chamber of Commerce. Unfortunately, I found that Cookiepedia data can categorise less than 22% of the cookies used by Alexa top20k websites and less than 15% of the cookies set

in the browsers of a set of real users. These results point to an acute problem with the coverage of current cookie categorisation techniques.

Consequently, I developed *CookieMonster*, a novel machine learning-driven framework which can categorise a cookie into one of the aforementioned four categories with more than 94% F1 score and less than 1.5 ms latency. I demonstrate the utility of the framework by classifying cookies in the wild. The investigation revealed that in Alexa top20k websites necessary and functional cookies constitute only 13.05% and 9.52% of all cookies respectively. I also apply the framework to quantify the effectiveness of tracking countermeasures such as privacy legislation and ad blockers. Results identify a way to significantly improve coverage of cookies classification today as well as identify new patterns in the usage of cookies in the wild.

6.1 Introduction

Browser cookies are ubiquitous in the web ecosystem today. Although these cookies were initially introduced to preserve user-specific states in browsers, they have now been used for numerous other purposes, including user profiling and tracking across multiple websites. This chapter sets out to understand and quantify the different uses for cookies, and in particular, the extent to which targeting and advertising, performance analytics and other uses which only serve the website and not the user add to overall cookie volumes.

First introduced in the mid-nineties as a way of recording client-side state [Net02], cookies have proliferated widely on the Web, and have become a fundamental part of the Web ecosystem. However, there is widespread concern that cookies are being abused to track and profile individuals online for commercial, analytical and various other purposes [SRDK+19a]. Recently, there has been a movement to restrict their usage, and companies such as Google have announced plans to replace certain kinds of cookies with more privacy-friendly equivalents [Bin21]. Before such drastic changes, however, it is important to take stock and understand how cookies are being used across the Web. Given the variety and number of uses for cookies and the fact that practically every website uses them, this is a herculean task.

This chapter is the first attempt to address this problem and catalogue cookies in the wild. Currently, the most commonly used classification in English language websites is the one proposed by the UK International Chamber of Commerce (UK ICC). The UK ICC catalogues cookies into four broad categories [ICC12]: *strictly necessary* cookies, which are essential for the website's function (e.g., logins, shopping carts); *performance* cookies, which collect analytical information to improve a website's performance; *functionality* cookies which remember user choices such as preferred language or location, allowing personalisation of the website to the user; and *targeting/advertising* cookies, typically placed by third-party advertising networks with the permission of the first party website to profile users and serve them ads.

The starting point is Cookiepedia, a database of over 31 Million cookies, which are categorised into the four UK ICC categories. Unfortunately, however, the measurements show that when queried with the cookies from the top20k websites according to Alexa[1], Cookiepedia can only identify and categorise around 22% of the cookies. I then turn to a Chrome plugin which I developed previously [HdTS20], and is currently being used by over six thousand users. 475 of these users (from 44 countries) are continuously donating anonymised cookie data to Thunderbeam[2]. Cookiepedia coverage on this dataset is even lower – it can classify less than 15% of this sample of cookies in the wild.

To address this problem, I treat the Cookiepedia data as a giant labelled dataset of cookie categories, using which I train a number of standard machine learning models, using a standard 5-fold cross-validation. Several of these models perform well, and I obtain a best-of-class F1 measure of around 0.95 with the Multinomial Naive Bayes classifier. All models rely on lexical n-gram features generated from the *names* of cookies. I then show that the model, which I term as *CookieMonster*, not only performs well in automatically categorising cookies found in the Cookiepedia

[1] https://alexa.com, which provides widely used ranks for websites

[2] This study is approved by the university ethics No. MRSP-19/20-18077

data, but also generalises to other cookies in-the-wild. I manually classify cookies on a random selection of Alexa Top 1 Million websites that are not in Cookiepedia, by leveraging GDPR consent managers used on these websites to allow users in the EU to decline particular categories of cookies. I demonstrate that the model is able to correctly predict (94% accuracy) the cookies which will be removed when a given category of cookies is declined through GDPR consent management, which indicates that models are able to correctly categorise cookies in-the-wild.

Inspired by this performance on websites not represented in Cookiepedia, I then use the model on all cookies in Alexa top20k websites, and find that the necessary and functional cookies (which are the two categories that are directly beneficial mainly to the user and not the website) constitute only 26.52% and 9.52% respectively of all cookies. Furthermore, I demonstrate for the first time that there are a number of third party cookies which are multi category. I then look at cookies donated by the users of the browser plugins, and find that even smaller percentages – less than 9.52% (respectively 13.05%) of cookies found in-the-wild are necessary (resply. functional). Interestingly, tracking/advertising cookies comprise 59.99% of cookies in the browsers of users from EU countries and a nearly similar 61.33% of cookies in non-EU countries, which is disturbing as it implies that EU users are not effectively utilising GDPR consent management to decrease the numbers of trackers in their browsers. I find similar results for other jurisdictions where there are web privacy-related laws, such as California (CCPA) or Brazil (LGPD). I also find that ad blockers are not fully effective, managing to block between 40–80% of all the third party advertising cookies.

6.2 *CookieMonster*: A system to understand cookie categories

In this section, I present the attempt to categorise cookies first using Cookiepedia [One20] and identify its inadequacy. Then I demonstrate how I designed *Cook-*

ieMonster using a data-driven approach to enable large-scale accurate cookie categorisation.

As I mentioned in section 6.1, I first attempted a simple off-the-shelf approach using Cookiepedia. Cookiepedia is an open-source database of browser cookies containing cookie details as well as their categorisation according to cookie usage. Cookiepedia is maintained by OneTrust, a privacy management software company and reports existence of 31,553,377 cookies [One20] in their database.

Cookiepedia provides a simple online search interface to search for cookie names. To that end, I first used browser automation using Selenium [Sel21] to collect all active cookies from `Alexa global top20k` websites. In total these globally most popular 20,000 websites used 54,694 unique cookies (with unique cookie names, i.e., cookie identifiers) for their visitors. In order to categorise these cookies, I query Cookiepedia with each of the cookie names using a Selenium-driven automated browser. For each of these cookies, Cookiepedia returned one of six categories: Strictly Necessary Cookies (essential for features of the website), Performance Cookies (used to collect information about how visitors use a website), Functionality Cookies (allow websites to remember user preferences), Targeting/Advertising Cookies (used to deliver personalised advertisements to users), Unknown and Nonexistent. The first four of these categories are based on UK ICC categorisation, which is also used in GDPR cookie consent management platforms [Col18]. An "Unknown" category indicates that the cookie exists in the Cookiepedia database but is not classified. A "Nonexistent" label indicates that a particular cookie does not exist in the Cookiepedia database.

According to the result of Cookiepedia-driven categorisation presented in §3.2.5. I make an surprising yet important observation–nearly 80% of the cookies used by Alexa top20k websites simply remain uncategorised when I use Cookiepedia database. Thus, even a massive database like Cookiepedia simply fell short in categorising the majority of the cookies used in even most popular websites today. To that end, in order to improve the categorisation of cookies while ensuring high accuracy and coverage I design and evaluate *CookieMonster*.

6.2.1 *CookieMonster* Design

The key idea of the system is to use machine-learning for accurate cookie categorisation in the wild. The ground truth for the classifier is the cookies collected from Alexa 20k websites which is classified in one of the four meaningful categories via Cookiepedia. There were 11,578 such cookies (Table 3.7 in §3.2.5) with their categorisation into four categories–Strictly Necessary, Functionality, Performance, Targeting/Advertising. For these cookies I used features extracted from the *cookie names* to train the classifier.

Next, I describe how I built the core classifier for *CookieMonster* using this ground truth data.

Preprocessing and tokenising cookie names

Each cookie is a name-value pair and the cookie-name is unique for each cookie. I noted via manual inspection that cookie names can be meaningful and appear to provide some hints about functionality. Thus I decided to use features extracted from these names for categorisation. First, I removed all numbers from each cookie name ((e.g., ADS_324 became ADS_). Next, I tokenise these names using punctuation characters (e.g., %, ~, ., _, -). Thus, at the end of preprocessing and tokenisation, a cookie with the name *gdpr-track-status45* will be split into tokens *"gdpr", "track", "status"*. Furthermore, I split the resultant token using capitalisation (i.e., AnalysisUserId → [Analysis, User, Id]) and used the enchant dictionary [Tho20] to segment known word combinations into root words (i.e, dayssincevisit → [days, since, visit]). Finally, I case-folded all the resulting tokens. In total, after this tokenisation, I retrieved a total of 2,504 unique tokens from 11,578 cookies in the ground truth data.

Manually checking correlation of cookie categories and tokens

Next, to verify the resultant tokens are meaningful, I divided the names into four cookie categories as provided by Cookiepedia. **I focused on the relationship between tokens and cookie categories.**

It could be noticed that some particular tokens and token combinations were immensely frequent in cookies from specific cookie categories. For example, cookie combinations like (gat, gtag) are popular within Targeting / Advertising cookie names. In fact, many popular tokens (e.g., geo, country, location, global) given by the cookie names in Targeting/advertising categories identify their usage in location tracking. Furthermore, names of third-party trackers are also frequent in these tokens (e.g., OWOX, Marketo, demdex). Although preliminary, the manual inspection of tokens gives the confidence that these tokens are correlated with cookie categories and using them as features in a supervised learning framework has the potential to be successful.

6.2.2 Supervised Cookie Categorisation in *CookieMonster*

Training a classifier for *CookieMonster*

Considering the labeled data would be pre-processed (cleaned, randomised, and structured) before training, and the cookie values are classified into more than two categories, I model the cookie categorisation as a supervised multi-class classification problem to predict four cookie categories—strictly necessary, functionality, performance and targeting/advertising. Given a cookie name, I extracted the tokens from the names (as mentioned above) and used them as features. Consequently, I evaluated seven classification algorithms to check the performance and identify which one to use in *CookieMonster*. I used the known categorises of cookies (from Cookipedia) as the training data. Specifically, I evaluated Multinomial Naive Bayes (MNB), Softmax Regression (Multi-layer perception or MLP), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forest, Naive Bayes and Binary Search Tree (BST). I used a 5-fold cross validation with 80-20 split between training and testing data. I used overall (Micro) precision, recall and F1-score over all-classes to report the accuracy of categorisation for all of the seven models in Table 8.3.

I make two observations from this table: First, the top four algorithms according to F1-score (MNB, MLP, SVM, KNN) all achieved F1-scores more than 0.9, signifying the

Algorithm	Precision	Recall	F1	Latency (ms)
Multinomial Naive Bayes (MNB)	0.951	0.940	0.9458	0.44
Softmax Regression (MLP)	0.944	0.948	0.9457	1.29
SVM **(linear kernel)**	0.947	0.867	0.926	0.03
K-Nearest Neighbors (KNN)	0.929	0.907	0.916	3.23
Random Forest	0.886	0.770	0.778	9.73
Naive Bayes	0.798	0.747	0.833	0.02
Binary Search Tree (BST)	0.649	0.461	0.409	0.05

Table 6.1: Recall, Precision and F-score of for different classification models to categorise cookies. MNB and MLP achieved more than 94% average F1-score, **which are relatively higher than other classifers**. Noted that "Latency" here represents for " Mean prediction Latency".

utility of the proposed features based on tokenising cookie names. Second, the top two algorithms (MNB and MLP) both achieved a F1-score of more than 0.94, making them suitable for use in *CookieMonster*. To that end, given I envision *CookieMonster* to be used in the wild for cookie categorisation, I next check the average categorisation latency for all of these classifiers.

Latency of prediction for classifiers

I present the average prediction latency for predicting the category of a single cookie during testing in Table 8.3. I note that, models like Bernoulli Naive Bayes, although extremely fast, provides a relatively poor F1-score (0.83). To that end, I focused on the top two classification models (MNB and MLP). These two models, while ensuring an F1-score of nearly 0.95, are quite different in terms of prediction latency. In fact, MLP has an average prediction latency of 1.2860 ms which is 293% higher than MNB. Therefore, I choose this pre-trained Multinomial Naive Bayes (MNB) model to use in *CookieMonster*.

Characterising Misclassified cookies in MNB classifier

I further did a simple analysis to understand why MNB model did misclassify a few cookies. I present the confusion matrix for MNB classifier from one fold of cross validation in Table 6.2. This shows that out of 2,016 cookies (the test set in this fold), 316 cookies got misclassified. However, the majority (244 out of 316) of this misclassification can be attributed to Necessary cookies being predicted as Targeting / Advertising and Targeting / Advertising cookies predicted as Performance cookies. I hypothesise two reasons for this. First, the Targeting/Advertising cookies share similar tokens with other cookie category. Second, Necessary and Performance cookies might sometimes also act as Targeting/Advertising cookies. I leave exploring these avenues to future work.

		Predicted				
		Nece.	Perf.	Func.	Target	Total
Actual	Nece.	486	1	2	120	609
	Perf.	2	566	7	16	591
	Func.	1	4	195	22	222
	Target	2	104	6	762	874
	Total	491	675	210	920	

Table 6.2: Confusion matrix of Multinomial Naive Bayes (MNB). Majority of the misclassification happened due to Targeting/Advertising cookies.

Finally, I note that overall (in spite of some misclassification), the accuracy of this fast MNB-based model is quite high in the training set (trained over from 11,578 cookies), however it makes a basic assumption—tokens extracted from a new cookie name will be included into 2,504 tokens that came from 11,578 cookies in the dataset. Clearly, this assumption might not hold in the wild cookie categorisation and I might encounter *out-of-vocabulary* tokens, which *CookieMonster* will need to address when used in-the-wild.

N-gram based additional categorisation for cookies with previously unseen tokens

New cookie names might contain tokens which are not in the list of 2,504 tokens seen in the training dataset of 11,578 cookies. Inability to categorise these cookies poses a challenge to the categorisation coverage of *CookieMonster*. This problem is common in NLP tasks which needs to deal with OOV (out-of-vocabulary) words (thus I call unseen tokens OOV tokens). **As previous collection has shown that most**

cookie names are semantically and functionally consistent, to solve this challenge, I designed an additional n-gram based classification for new cookies.

In the proposed approach, a new cookie name (e.g., *_bti*) with previously unseen tokens is simply divided into the constituent character n-grams (e.g., *_bti* can be split into bi-grams [('b', 't'), ('t', 'i')]). In the Cookiepedia dataset I noted that 75% of cookie names have 5 or less characters. So I choose to use $n = 2, 3$ and 4. Next I simply search for these n-grams within the set of the 11,578 cookie names and create a set of existing cookie names that contain these n-grams (e.g., *NSC_mc-vsmibti* and *gati_abtc* which matched bigram of *_bti*). Finally, out of these existing cookie names I choose the one with the least edit distance with the new cookie name and output the category of that existing cookie as predicted category of the new cookie. In the example, since edit_distance (_bti, NSC_mc-vsmibti) = 10 and edit_distance (_bti, gati_abtc) = 6, so I predict category of *_bti* to be the same as the category of *gati_abtc*.

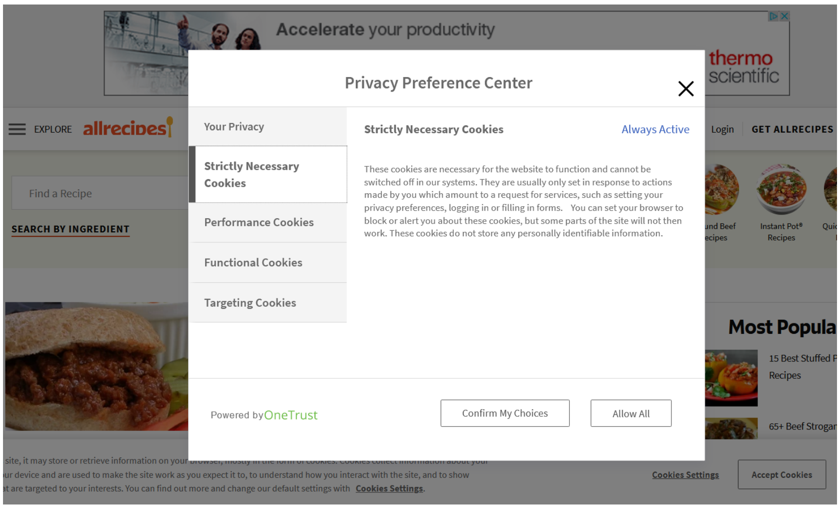
Final workflow of *CookieMonster*

So, to summarize, *CookieMonster* used cookie names to categorise cookies. On encountering a cookie name, *CookieMonster* will run the pre-processing step and identify tokens from the cookie names. If those tokens exist in the MNB-based pretrained model, then *CookieMonster* will output the prediction of MNB classifier. Otherwise, it will use the ngram based additional classifier to find a previously seen token that is lexically similar to the new unseen token, and will predict the cookie category based on the known tokens. However, one obvious question is: *since CookieMonster primarily uses the Cookiepedia data for its design, can it accurately classify cookies in-the-wild on websites not catalogued in Cookiepedia?* I answer this question affirmatively in the next section.

6.3 Cookie categorisation in-the-wild

CookieMonster gives a tool to examine a collection of cookies and categorise them into the 4 widely used UK ICC categories. I first perform a manual verification (§6.3.1) on websites *not* included in Cookiepedia, to show that *CookieMonster* generalises widely. Then, given that I have a reasonably accurate method to classify cookies beyond the dataset it is trained and tested on, I ask what proportion of cookies are superfluous to a user's experience of websites, looking both at the top20k websites according to Alexa, and at cookies found in browsers of real users in-the-wild (§6.3.2). Finally, I use *CookieMonster* to quantify the effectiveness of current web privacy measures (§6.4).

6.3.1 Does *CookieMonster* work in-the-wild? – a manual verification



Strictly Necessary Cookies

These cookies are necessary for the website to function and cannot be switched off in our systems. They are usually only set in response to actions made by you which amount to a request for services, such as setting your privacy preferences, logging in or filling in forms. You can set your browser to block or alert you about these cookies, but some parts of the site will not then work. These cookies do not store any personally identifiable information.

site, it may store or retrieve information on your
ur device and are used to make the site work as you expect it to, to understand how you interact with the site, and to show
at are targeted to your interests. You can find out more and change our default settings with Cookies Settings.

Figure 6.1: Cookie Consent Example

section 6.2 demonstrated that cookie names can reveal the purpose and UK ICC category of the cookies. While this was rigorously tested using 5-fold cross validation on Cookiepedia data, I still need to validate whether the model can correctly identify the purpose of cookies on websites which have *not* been catalogued on Cookiepedia.

This is not straightforward, as the *purpose* of cookies on most websites may not be apparent.

To answer this question, I take advantage of GDPR, which holds in the European Union (and in the UK vantage point). GDPR requires websites to obtain user consent before collecting data about them. Because of this, it is extremely common to see websites using consent management banners such as the example shown in Figure 6.1. As in the figure, many websites use the UK ICC categories for allowing users to control their consents. Thus, a careful user can control which categories of cookies are allowed from a given website. With the website in Figure 6.1, users *have* to allow necessary cookies (there is no choice), but may choose to allow additional categories of cookies. For example, one user may decide to allow necessary and functional cookies. Another user may allow necessary and performance cookies instead. Clearly other combinations are also possible, including allowing three or all four categories of cookies. This is a common pattern for consent management in many websites.

I can therefore determine which cookies are in the "necessary" category by visiting the website with a clean browser (after deleting all cookies and clearing the user profile) and selecting to allow only the necessary cookies. I can then clear the user's cookie and profile information again and revisit the website, this time choosing to allow necessary *and* functional cookies. The *additional* cookies installed in this second visit can be inferred to be in the "functional" category. A similar approach can be used to determine "performance" and "advertising/targeting" cookies.

The above approach is not scalable, but serves to test whether the *CookieMonster* model "works" beyond the Cookiepedia data. To this end, I select websites that satisfy two criteria: *(i)* They are *not* indexed in Cookiepedia (to test generalisability of the model). *(ii)* They have deployed a GDPR consent management solution that allows free choice among the four UK ICC categories (so that the approach above can be applied on that site). I randomly select $n = 60$ websites satisfying the criteria, choosing 10 each from the Alexa 1-100, 101-500, 500-1000, 1K-10K,, 10K-100K and 100K-1M ranks. I note that much of the Cookiepedia data comes from a database maintained by OneTrust[3]. Among the 60 sites I choose, 7 sites do use OneTrust (Table 6.3), although these sites are still not indexed in Cookiepedia. Thus, the manual test verifies generalisability beyond Cookiepedia data to sites with and without OneTrust support.

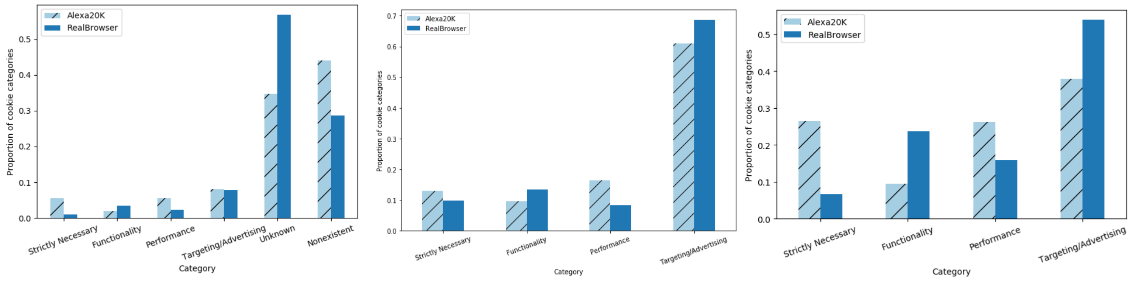
	Recall	Precision	F1-score	OneTrust	OOV(%)
top1-100	0.93	0.87	0.91	2	0
top100-500	0.90	0.86	0.88	3	0.83%
top500-1k	0.83	0.85	0.87	0	1.68%
top1k-10k	0.86	0.93	0.89	2	0
top10k-100k	0.79	0.77	0.78	0	4.61%
top100k-1M	0.74	0.84	0.79	0	6.17%

Table 6.3: Recall, Precision and F1-score of *CookieMonster* for cookie recognition across `Alexa top-1M` websites. OOV is the percentage of cookies which were not recognised and had to be classified using the OOV technique (§6.2.2). The OneTrust column identifies the number of websites in each category using OneTrust GDPR Consent Management.

Table 6.3 shows that the model generalises extremely well. As may be expected, the performance is best for the top ranked Alexa sites (F1 score > 0.85 for the Top10K sites), but even in less popular sites up to Alexa rank 1 Million, an F1-score of > 0.78 is obtained. For each category of ranks, I also show the proportion of cookies whose names contained previously unseen tokens and therefore required the OOV technique (§6.2.2) to be used. Most cookies are recognised within the model and OOV matching is required for less than 6-7% or fewer cookies.

I conjecture that *CookieMonster* generalises beyond the Cookiepedia data it is trained on because it is based on cookie *names*, which are set by the JavaScript libraries or the third party providers a website uses for targeting, advertising, analytics etc. The choice of a website to use a particular GDPR consent management platform such as OneTrust (which impacts inclusion in the Cookiepedia database) is orthogonal to the libraries and third party providers (and therefore the cookie names) it uses. A few libraries and third party providers dominate the ecosystem in each country [HdTS20]; thus cookie names or the naming pattern n-grams used in *CookieMonster* generalise across websites.

[3] https://cookiepedia.co.uk/about-cookiepedia



(a) Cookiepedia

(b) *CookieMonster*

(c) Cookiepedia (ignoring the unrecognised and uncategorised cookie names)

Figure 6.2: Proportions of different cookie categories in Alexa top20k (shaded) and real browsers (clear), according to (a) Cookiepedia (b) *CookieMonster* (c) Cookiepedia (ignoring the unrecognised and uncategorised cookie names)

6.3.2 What proportion of cookies are actually required for websites to function properly?

Strictly speaking, a user only needs to enable “necessary” cookies (e.g., login or shopping cart cookies). Some may choose enable “functionality” cookies that personalise a site (e.g., to user’s preferred language or site layout). Arguably, performance analytics and advertising/targeting cookies benefit the website more than they do the user and do not need to be enabled. *CookieMonster* therefore provides a convenient way to quantify how many cookies are superfluous.

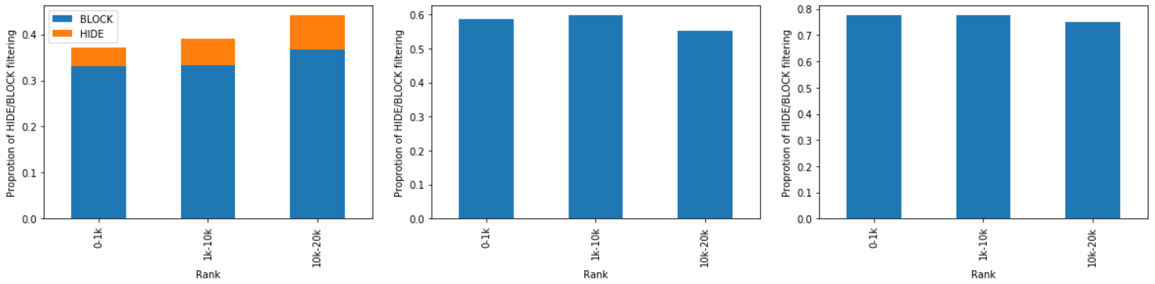
I study this systematically in Figure 6.2, by categorising all the cookies of the Alexa top20k websites as well as cookies collected from users of a browser extension I developed and deployed in an earlier study [HdTS20], and is currently being used by over 6000 users. Specifically, in this work I use 44,971 cookies collected between November 2020 to February 2021 from 475 of these users (from 44 countries) who are donating their data. I use two methods for the categorisation: looking up the cookie name in the Cookiepedia database (Figure 6.2(a), which presents the same information as Table 3.7), and using *CookieMonster* (Figure 6.2(b)) to predict a category. As mentioned previously (cf. subsection 3.2.5), the Cookiepedia database is fairly incomplete, with over 78% of cookie names either not existing in the database or not categorised; thus, for the purpose of comparing with *CookieMonster*, I replot Figure 6.2(a) by ignoring these unrecognised and uncategorised cookies and renormalising the remaining cookies as 100%, obtaining Figure 6.2(c).

Both Cookiepedia (Figures 6.2(a), 6.2(c)) as well as *CookieMonster* (Figure 6.2(b)) show similar trends: According to *CookieMonster*, only 13.05% of cookies are labeled as necessary, and an additional 9.52% are functional. According to Cookiepedia, 5.6% of cookies are labeled as necessary (26.52% after ignoring unrecognised/uncategorised cookies), and an additional 2.01% are functional (9.52% after ignoring unrecognised/uncategorised). Thus, both methods suggest that *the vast majority of cookies can be removed without affecting user experience.*

Interestingly, according to both *CookieMonster* (Figure 6.2(b)) and Cookiepedia (Figures 6.2(a), 6.2(c)), real browsers have a smaller proportion of necessary cookies and more functional/targeting cookies as compared to Alexa top20k websites. This is likely because real users' browsers have user profiles which are better established, with a browsing history and long-lived cookies that may have been set months ago, leading to better profiling and more ads/targeting cookies. In contrast, I collect cookies on Alexa top20k websites programmatically using Selenium with a fresh user profile instance for each website, resulting in fewer ad/targeting cookies. Also, the user base is located in different countries where there may be country-specific third party trackers [HdTS20] not visible from the UK vantage point, and therefore not captured in the Alexa crawl.

6.4 On the effectiveness of current web privacy measures

The previous section suggests that a large proportion of cookies can be eliminated from many websites without affecting their function. One of the main levers of control that users can employ to achieve this, is to use ad blockers. In addition, web privacy regulations around the world, such as GDPR, provide varying degrees of support



(a) EasyList+Alexa Topsites (b) EasyPrivacy+Alexa Topsites (c) AdGuard+Alexa Topsites

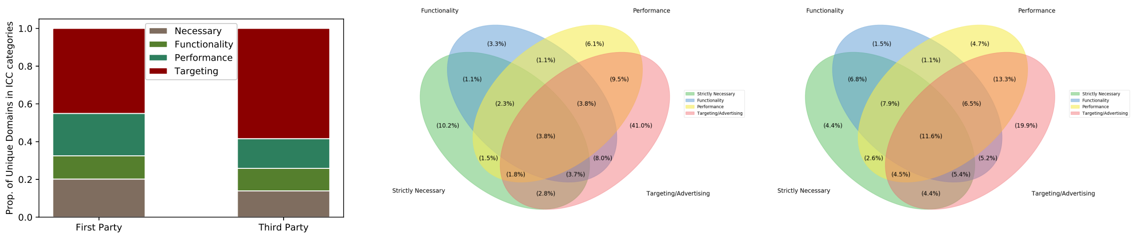
Figure 6.3: EasyList, EasyPrivacy and AdGuard filter 40–80% of advertising third party cookies on Alexa top20k sites.

for users to provide consent or decline different kinds of cookies. I examine their effectiveness below.

6.4.1 Ad blockers

Ad blockers typically work based on dynamically updated lists of third party advertising/targeting domains that should be blocked. Figure 6.3 shows how three popular block lists – EasyList, EasyPrivacy and AdGuard Plus – work on cookies found in Alexa top20k websites. In addition to a block list, EasyList has a so-called ‘hide’ list of domains which break if blocked, and therefore, are loaded but not rendered on screen, to improve user experience. Unfortunately, because the domain is loaded, the user can still be tracked even if the ad itself is hidden. These domains are therefore shown separately. In general, Ad Guard appears to block a larger proportion of domains than EasyList or EasyPrivacy, even when counting cookies from hidden domains in addition to the cookies from blocked domains. I also find that there are more domains to be blocked in real browsers than when visiting Alexa top20k sites programmatically. Again, this is likely because of additional targeting and advertising that may tend to be attracted by more mature user profiles with a continuous browsing history.

Across all the combinations tested in Figure 6.3, I still find that around 20% (for Ad Guard Plus) to 60% (for EasyList) of advertising and targeting-related cookies that should have been blocked are not being blocked. This is partly because the lists that



(a) Relative proportions of different ICC categories in FP and TP cookies

(b) Shared (TP)

(c) Shared (FP)

Figure 6.4: Occurrence of multipurpose Third-Party domains in top20k websites.

ad blockers rely can never be complete. However, when I dig deeper, I on find two additional important reasons: First, ad blockers are relatively successful at blocking third party advertising and cookies, but I find that a significant proportion of first party cookies also relate to advertising. Figure 6.4(a) quantifies this, showing the relative proportion of targeting cookies and other categories of cookies among both first party cookies and third party cookies. Thus, several first party cookies may slip through ad blockers. Secondly, I find that both among first parties (Figure 6.4(c)) and third parties (Figure 6.4(b)), a non-trivial proportion of advertising-related domains also place other categories of cookies. Thus, a solely domain-based block list risks either blocking too much, or not covering all the domains that undertake targeting. The domain-based approach is common among all widely used ad blockers – the diversity of cookies on the web has thus far made it difficult to take a more granular approach that blocks specific cookies. However, since *CookieMonster* appears to provide reasonable predictions of cookie categories based on cookie names, I may use it as one component of a more sophisticated system that blocks specific cookies. Such approaches can complement other methods which have utilised the Internet Advertising Bureau's *Ads.txt* [EJRHPAF19] and other list-based measures to identify ads.

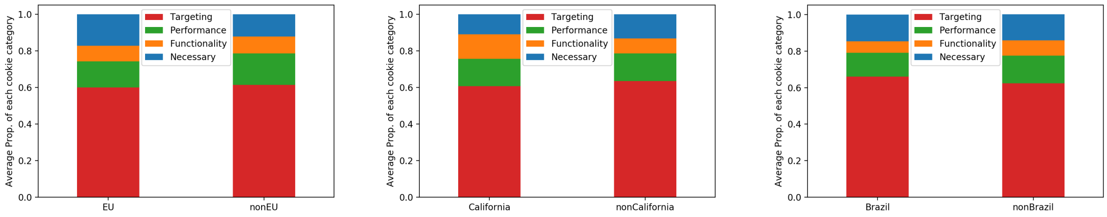
6.4.2 Privacy regulation

A second lever that users have recently obtained is support from privacy-related regulations in various legal jurisdictions. By far the most comprehensive and well-known of these is the General Data Protection Regulation (GDPR) in the EU, which introduced the notion of requiring explicit and meaningful consent. Comparable regulations include the California Consumer Privacy Act (CCPA) which allows users to opt-out of tracking and Brazil's Lei Geral de Proteção de Dados (LGPD), which is the most recent of them, and also mandates unambiguous consent from users before websites can use cookies.

Previously, using a limited cohort of 16 users, I had found that cookie numbers seen by users had not changed significantly before and after GDPR was introduced [HS19], implying that users may be choosing the 'default' choices offered by websites, which may not be privacy optimal. Here, I extend this study based on the 475 users of the extension [HdTS20] who are donating data. Specifically, I consider all users within a given privacy jurisdiction (EU, California or Brazil) and compare the proportions of ad/targeting cookies of users from within that jurisdiction to the respective proportions in browsers of users outside the jurisdiction. Figure 6.5 shows that in all cases, there is little difference between proportions of cookies of users within and out of each of the jurisdictions. This confirms (using a much larger user base) the previous finding [HS19] that users are not making the most privacy optimal choices for themselves, and may be fatigued the burden of providing consent on every website they visit, especially as several websites use dark patterns that make it difficult to choose more privacy-oriented settings [NLV^{+}20].

6.5 Discussion

This chapter set out to tackle the herculean task of classifying cookies found in-the-wild. I started with data curated on Cookiepedia, and demonstrated that its coverage was inadequate – its database contained less than 22% of cookies on Alexa top20k



(a) GDPR: EU vs. non-EU

(b) CCPA: California vs. non-California

(c) LGPD: Brazil vs. non-Brazil

Figure 6.5: Proportions of ad/targeting cookies within and outside of 4 jurisdictions with privacy regulations.

websites, and less than 15% of cookies found in real browsers. I therefore developed machine learning models that trained on Cookiepedia data and were also shown to work well ($F1 > 0.94$) on websites not currently in Cookiepedia. Models use lexical features derived from cookie names, suggesting that cookie names generalise well across websites, perhaps as a result of common web templating infrastructures and libraries, and the prevalence of common third parties across websites.

Then, I used the trained models on the Alexa top20k websites as well as the anonymous cookies donated to Thunderbeam by 475 users of a plugin I have developed previously [HdTS20]. I found that across the 44 countries represented in the dataset, necessary and functional cookies (the two categories beneficial to the user rather than the website) constitute only 9.79% and 13.35% of all cookies in the active countries. Thus, the vast majority of cookies can be removed without impacting website functionality or user experience.

Surprisingly I find that privacy regulations such as GDPR in the EU have not made much difference in the numbers of cookies seen by real users. This indicates that users are not effectively utilising the consent management options enabled by GDPR. Ad blockers appear to be more effective if used, but mainly focus on advertising cookies. Even among advertising cookies, a non-trivial proportion is missed because the ad blockers are based on *manually curated* lists [Ato05] which need to be continuously updated and because these lists are based on blocking at the level of the domains that serve up those cookies, rather than on blocking specific cookies. Unfortunately, I also find that many domains set both non-essential (e.g., advertising or performance) as well as essential (necessary or functional) cookies; thus extreme care needs to be exercised in blocking of entire domains, to ensure that functionality of the website is not broken as a result.

Thus far, the diversity of cookie names has prevented a more fine-grained approach and continuously updated but manually curated lists of domains to block have been the main tool for actively restricting tracking and cookies via ad blockers. I propose that the robust *CookieMonster* model based on lexical tokens extracted from cookie names can be used as the basis for sophisticated tools enable *automatic* rejection of specific cookies belonging to categories that are not beneficial for users. I intend to develop this idea in future work.

Chapter 7

Regulations and governance: GDPR

It is time to stop the anarchy on the Interne.

—— Alexander Lukashenko

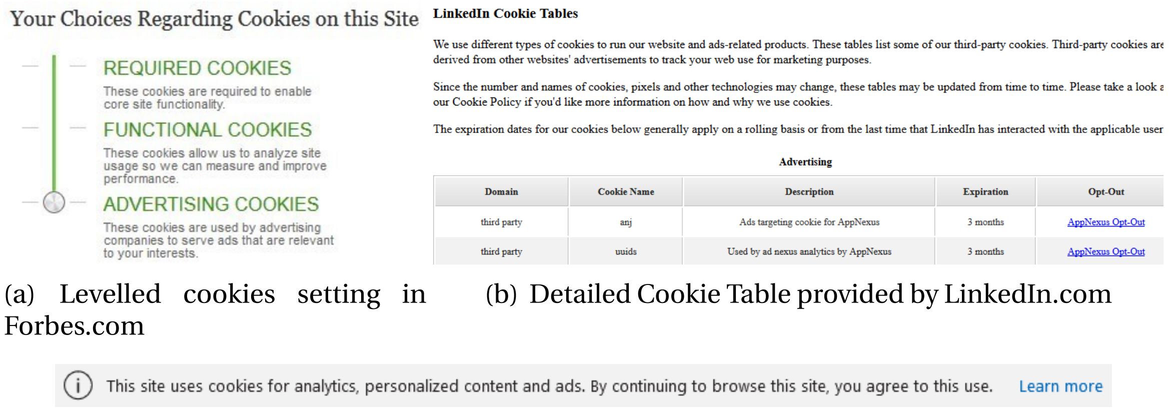
Summary

The recently introduced General Data Protection Regulation (GDPR) requires that when obtaining information online that could be used to identify individuals, their consent must be obtained. Among other things, this affects many common forms of cookies, and users in the EU have been presented with notices asking for their approvals for data collection. This chapter examines the prevalence of third party cookies before and after GDPR by using two datasets: accesses to top 500 websites according to `Alexa.com`, and weekly data of cookies placed in users' browsers by websites accessed by 16 UK and China users across one year.

I find that on average the number of third parties dropped by more than 10% after GDPR, but when I examine real users' browsing histories over a year, I find that there is no material reduction in long-term numbers of third party cookies, suggesting that users are not making use of the choices offered by GDPR for increased privacy. Also, among websites that offer users a choice in whether and how they are tracked, accepting the default choices typically ends up storing more cookies on average than

on websites that provide a notice of cookies stored but without giving users a choice of which cookies, or those that do not provide a cookie notice at all. I also find that top non-EU websites have fewer cookie notices, suggesting higher levels of tracking when visiting international sites. Findings have deep implications both for understanding compliance with GDPR as well as understanding the evolution of tracking on the web.

7.1 Introduction



Your Choices Regarding Cookies on this Site

REQUIRED COOKIES

These cookies are required to enable core site functionality.

FUNCTIONAL COOKIES

These cookies allow us to analyze site usage so we can measure and improve performance.

ADVERTISING COOKIES

These cookies are used by advertising companies to serve ads that are relevant to your interests.

(a) Levelled cookies setting in Forbes.com

LinkedIn Cookie Tables

We use different types of cookies to run our website and ads-related products. These tables list some of our third-party cookies. Third-party cookies are derived from other websites' advertisements to track your web use for marketing purposes.

Since the number and names of cookies, pixels and other technologies may change, these tables may be updated from time to time. Please take a look a our Cookie Policy if you'd like more information on how and why we use cookies.

The expiration dates for our cookies below generally apply on a rolling basis or from the last time that LinkedIn has interacted with the applicable user

Advertising

Domain	Cookie Name	Description	Expiration	Opt-Out
third party	anj	Ads targeting cookie for AppNexus	3 months	AppNexus Opt-Out
third party	uuids	Used by ad nexus analytics by AppNexus	3 months	AppNexus Opt-Out

(b) Detailed Cookie Table provided by LinkedIn.com

ⓘ This site uses cookies for analytics, personalized content and ads. By continuing to browse this site, you agree to this use. Learn more

(c) Office.com provides a cookie notice but no choice

Figure 7.1: Examples of cookie notices provided by website owners to EU users after GDPR came into effect (May 25, 2018)

The General Data Protection Regulation (GDPR) is a sweeping regulation that came into effect on May 25, 2018 in the European Union (EU), to protect the online privacy of its residents [Por18]. GDPR affects many aspects of personal data collection [TPRM18], although some argue that it does not go nearly far enough [Zar16b]. One of GDPR's central tenets is that whenever personal data is collected about a user, it has to be done with the consent of the user.

This notion of user consents has affected a large number of sites that have used various mechanisms including analytics, tracking, and targeted advertising to track users. Such websites are now required to inform users. Consent for cookies which can be used to identify a user uniquely is explicitly mentioned in Recital 30 [3018].

The need to inform users has led to a large number of cookie notices to users. Different websites have adopted different practices as shown in Fig. 7.1. Some, such as Forbes and LinkedIn (Fig. 7.1 (a) & (b)) have provided users with several choices, allowing them to select or unselect different options. Others, such as Office.com (Fig. 7.1 (c)) simply inform (without giving the user any choice) that user-specific cookies are being used, and this notice needs to be accepted if the website is accessed. The last option is not to issue any notice at all, because either no user-specific cookie has been used, or non-compliance of GDPR. Many websites appear to have chosen one of the first two options (cookie notice with or without choice) because GDPR non-compliance can attract fines of up to the higher of 20 Million Euros or 4% of the turnover of a company[1].

In this chapter, I investigate GDPR cookie notices on two sets of websites. The first is the set of top sites according to Alexa Web Traffic Analysis. The second set comprises websites visited by real users in an ongoing study[2]. In both cases, I focus on so-called *third party cookies*, i.e., cookies set not by the "first party" sites visited by the users, but by other third parties used by the first party sites. For example, if a user visits a site that uses Google Analytics, a Google (Analytics) cookie is placed in the user's browser. Third party sites hold enormous power since they obtain a panoramic view of a user's browsing history across different sites using the same third party.

I access these sets of websites from a vantage point in the EU, and obtain the following results:

1. Generally, websites which offer users a choice store *more* third-party cookies (when users accept default options offered), than sites which do not give users a choice. Some websites appear to continue placing cookies that are used to track users even after they explicitly decline consent[3].

[1] Art. 83(4) and 84(5) of the GDPR. https://gdpr-info.eu/art-83-gdpr/

[2] All collected data have been obtained with agreement from participants and under Research Ethics Minimal Risk Registration process at the university to ensure the permissions of approvals relevant to this research (Ethics approval no. MRS-1718-6539)

[3] Example screencast videos for such websites in Top500: https://bit.ly/2GnWrim

2. The number of third party cookies, as well as the manner of GDPR consent notices, vary across different categories of websites. Adult websites are the least likely to offer GDPR consent and choices, but also appear to contain fewer third party cookies, likely because several common third parties such as Facebook and DoubleClick do not work with adult sites. In contrast, news websites have the highest number of third parties, and also provide more cookie consent notices.

3. The prevalence of third-party cookies appears to differ across countries: Nearly 90% (66%) websites in the `Alexa.com` Top 100 in China (USA) do not issue any third party cookie notices, or provide no choice to users on the manner of tracking.

4. On average, the number of third-party cookies from UK websites drops by 10% after May 25, 2018, suggesting that GDPR has been successful and sites are complying with the regulation. However, this reduction appears to not be reflected in real users' browsing histories, and third party cookie numbers in 2019 show little change since before GDPR.

7.2 Datasets

Results are based on two datasets. The first dataset focuses on the top websites, i.e., those which obtain the maximum amount of traffic according to `Alexa.com` [Ale18]. I first analysed the top 100 sites in the UK one week before and one week after the introduction of GDPR (May 25 2018), focusing on differences in cookie numbers. In addition, I manually examine the types of cookie notices served by the top 500 websites in the UK after GDPR has been introduced.

The second dataset is obtained from a study in which anonymised browser histories are being collected weekly from 15 users (9 in the UK; 6 in China). I have instrumented the browsers of these users using a modified version of a browser plugin, Lightbeam [Moz12] which runs also on Google Chrome. The plugins collect information about both the first party websites they visit, as well as the third party cookies placed as a result of visiting those first party sites. Altogether these users have visited around 15k first-party websites across the year, which have led to over 187k third-party domains from which cookies are placed on their computers (Table 7.1). I focus on the UK users who have visited around 8416 websites and have cookies from nearly 113K third-party domains.

User Group	**No. 1^{st} party sites**	**No. 3^{rd} party cookies**
UK Users	8416	113,003
CN Users	6144	74,313
Total	14827	187,316

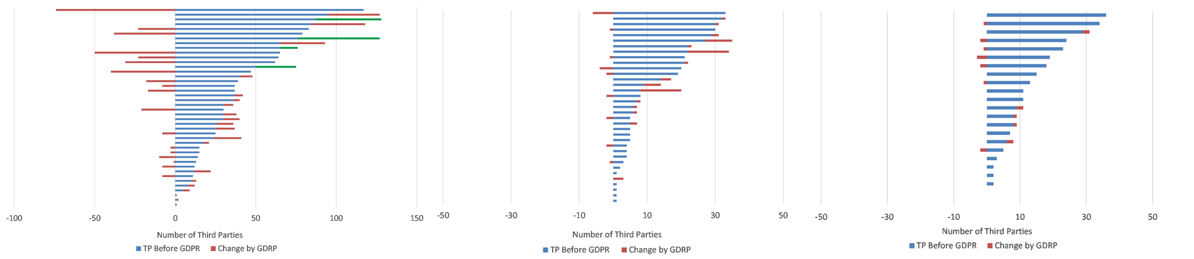
Table 7.1: Data collected from Jan. 2018 to Jan. 2019

7.3 GDPR notices in Alexa top websites

I first study GDPR cookie notices in popular websites. This study comprises three steps. **Since The first data collection was completed one week before the GDPR was enacted, to analyse the execuation of GDPR,** I capture cookies one week before and one week after GDPR comes into effect, among the `Alexa.com` Top 100 sites in the UK, which, as a current member of the EU, is subject to GDPR. Next I compare UK cookie notices after GDPR was introduced, with those from outside the EU, taking USA and China as examplar non-EU countries, and also using `Alexa.com`'s global lists of top sites in various important categories of the web, such as shopping and technology. I then manually examine the different kinds of cookie notices among the top 500 websites in the UK, and discuss the impact on tracking and GDPR compliance.

7.3.1 Cookie notices among Alexa Top 100 sites

After 25th May in 2018, websites started to pop up cookie notices to users before data from them is collected. Generally, there are three types of cookie notices: The



(a) Cookie notice with choice (42 sites)

(b) Cookie notice no choice (35 sites)

(c) No cookie notice (23 sites)

Figure 7.2: The changes on the number of third-party cookies of Alexa Top100 Websites (one week before and after GDPR), if the default choice is accepted. Each horizontal line denotes a site, totally 100 lines across three subgraphs. For each site, blue shows the number of third-party cookies served before GDPR, and red the *change* in the number of cookies after GDPR. Three categories are observed: (a) Sites which serve users with cookie notices. (Green indicate sites which store cookies even if users explicitly opt out) (b) Sites which serve cookie notices but offer no choice to users. (c) Sites which serve no notices after GDPR.

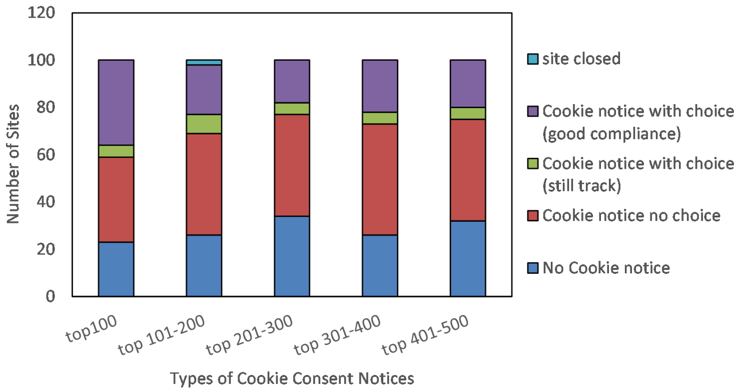


Figure 7.3: Detailed study of UK top500 sites' Cookie Types.

first one is that the website owner provides users with a privacy choice of opting out from the data sharing, e.g., Forbes and LinkedIn (Fig. 7.1 (a) & (b)). Other examples include Reddit, Twitter and Amazon.

The second kind of websites includes vendors that provide a notice of cookie collection but they do not offer a way to change the setting, e.g., Office.com (Fig. 7.1 (c)). Essentially, the user has to choose between using the website with cookies being used, and not using the website at all. The final kind of websites provide no cookie collection notice. A handful of websites also stop their business and support for European users. This includes several prominent non-EU sites such as LAtimes.com, ChicagoTribune.com, QQ.com, Unroll.me, etc.

Fig. 7.2 studies GDPR cookie notices of the `Alexa.com` Top 100 websites in the UK. Nearly 80% of these sites display some form of cookie notice (Fig. 7.2 (a) & (b)), and half of all collected websites provide an option on whether to receive personalised ads or not (Fig. 7.2 (a)). When the websites provide a choice, I accept the default settings and observe the number of cookies stored[4]. 22 websites in the top 100 do not serve any cookie notice.

As expected, GDPR appears to have had an effect on the *number* of third party cookies immediately after the law came into effect. Amongst websites which allow users to set their choices (Fig. 7.2 (a)), the average number of third party cookies dropped from 34 to 28; websites which show a cookie notice but provide no choice in the matter (Fig. 7.2 (b)) show a minor reduction from 16 cookies on average before GDPR to 15 after; those which do not issue cookie notices (Fig. 7.2 (c)) show no change, with an average of 13 third party cookies before and after GDPR.

Degree of GDPR compliance: It is interesting and notable that websites which appear to be transparent and offer users a choice (Fig. 7.2 (a)) store *more* cookies (avg. 28) when the default option is accepted, than those which provide no choice (avg. 15). Similarly, several websites which offer an option seem to have used the opportunity to *increase* the number of third-party cookies (Red lines on the positive side of Fig. 7.2 (a)). Examining manually, I see that websites which do not serve cookie notices use some of the same third party trackers (e.g., Google Analytics or Facebook cookies) which are found among websites that do serve notices, which suggests that perhaps such websites *should* be serving cookie notices and asking for user consent.

Furthermore, in the manual examination of websites that do provide users with a choice, I see cases where tracking cookies are being placed even after opting out of tracking and personalisation (i.e., even when I choose non-default choices that maximise privacy), highly indicative of GDPR non-compliance (See footnote 4). Fig. 7.2 (a)

[4] Note that some of the cookies stored are simply to note the fact that the cookie notice has been served and accepted. I discard these cookies from counts.

shows these websites with green, and it is interesting to note that these websites have higher than average number of cookies among those that provide cookie notices with choice.

Finally, Fig.7.3 expands this study from the top 100 sites I have been looking at so far to the `Alexa.com` top 500 sites. As expected, the fraction of sites offering users a choice drops drastically after the top 100. Many sites also close and stop serving EU users.

7.3.2 Cookie notices of top non-EU websites

GDPR compliance is a requirement for all websites that wish to operate within or can be accessed from EU locations. Therefore, I am interested in understanding how non-EU websites have dealt with the introduction of GDPR as they will also be subject to the regulation if serving EU citizens in the EU. As mentioned previously, several prominent websites such as LATimes.com (`Alexa.com` rank 163 in the USA), Chicagotribune.com (`Alexa.com` USA rank 342) and QQ.com (`Alexa.com` rank 2 in China), have once stopped serving users in the EU, serving up a banner that says they do not operate within EU boundaries because of GDPR.

Therefore, as a baseline, I manually examine how `Alexa.com` top 100 sites in China and the USA serve cookie notices when accessed from the UK. Table 7.2 shows the comparison of top 100 sites in the UK (also studied in Fig. 7.2) and those in China (CN) and USA (US). In contrast with the UK, only 10% (respectively 34%) of sites in China (USA) offer users a choice of which cookies to store, and only a further 6% (14%) serve a cookie notice with no choice. Thus the vast majority (84% in CN, 52% in the US) of top sites are currently operating without a cookie notice. A large proportion also serve a notice that tracking cookies are being used, but users are not able to opt out of such cookies and continue to use the websites. Indeed, only a small fraction 10% (34%) of top sites in CN (US) offer users a cookie notice with choice. Therefore, it appears that *users of international non-EU websites in the UK obtain little protection, and little choice about their privacy and tracking.*

Types of GDPR notices in sites	UK	US	CN
Cookie notice with customised options	42%	34%	10%
Cookie notice but no customised options	35%	14%	6%
No cookie notice	23%	52%	84%

Table 7.2: Percentage of different types of cookie notifications on Alexa Top 100 websites from the United States (US), China (CN), and the United Kingdom (UK) within one year after the GDPR was released (2018-2019).

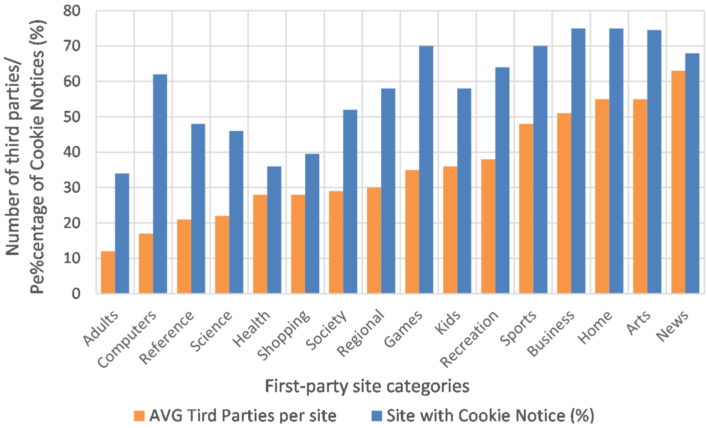


Figure 7.4: The average number of third parties per site and percent of cookie notices in each category.

I next turn to global top sites across categories in `Alexa.com`, to understand GDPR compliance among different kinds of websites. Fig. 7.4 shows the categories ranked by the number of third parties per site for each category on average. The count in Adult websites is the least, likely because they typically are not able to access the most common third-party cookie providers such as Facebook or Google Analytics. However, Adult websites also have the lowest fraction of websites serving cookie notices. News and home related websites have the largest number of third parties, but also show the highest levels of compliance (i.e., serve cookie notices). In general however, no individual category of global websites achieves the same level of compliance as the top 100 UK websites.

7.4 Cookie notices to real users

Until now, I have been studying how top sites around the web serve third-party cookie notices. However, any given user may have niche interests, and will likely access sites outside the list of `Alexa.com` top sites. To understand how compliant those less popular sites are, I turn to an ongoing user study I am conducting on third-party trackers collected by browser plugins, using a live user group. I also wish to understand whether real users see a decrease in number of tracking third party cookies after GDPR.

Cookie notices in real users' browsing histories I use 1528 websites collected by UK users in the weeks from Jan - Mar 2018 and evaluate the popularity of those sites by their visiting frequency to group them into 5 quintiles. **Quintiles are used to create cut-off points for a given population in a socio-economic stud, here I make use of it to explore the relationship between the cookie banner and the website popularity.** Quintile 1 comprises 133 sites visited by over 80% participants, quintile 2 has 150 sites visited by around 60% - 80% users, quintile 3 is 148 sites visited by 40% - 60% users, quintile 4 168 sites by 20% - 40% users and quintile 5 has by far the most number of sites (929), but each site is visited by less than 20% participants. Even the `Alexa.com` top 100 sites are evenly distributed across the five quintiles – 15 of the `Alexa.com` UK top 100 sites fall in quintile 5, i.e., are visited by fewer than 20% of users. 19 `Alexa.com` top100 sites are not accessed by *any* user.

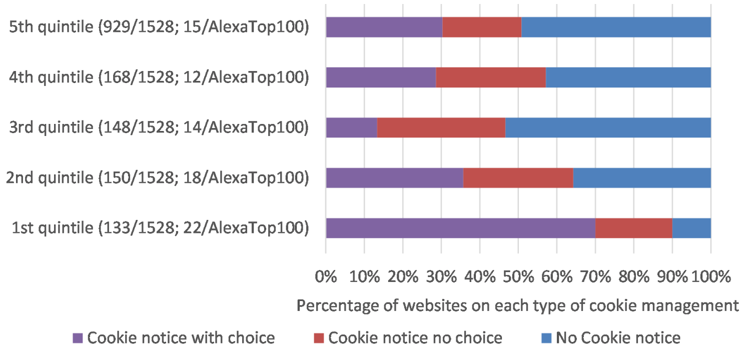


Figure 7.5: Cookie notices among the five quintiles of websites accessed by a real user base **that collected from Jan - Mar 2018**.

Fig. 7.5 shows the distribution of different kinds of cookie notices among the websites in different quintiles. Reassuringly, websites which are visited by most of the users in the study (quintile 1) has the highest fraction of websites which serve some form of cookie notice. However, as I go towards more niche interest websites, those visited by smaller numbers in this user study, the fraction that serve GDPR cookie notices drops drastically (there is a steady decline up to quintile 3, and although there is a brief uptick in quintiles 4 and 5, the fraction serving cookie notices are still below the top 2 quintiles). This suggests that users may need to be careful about niche websites.

Did GDPR affect third party cookie numbers for real users? Whereas previous sections have looked at synthetic or programmatically generated browser visits to websites, I can also ask the *extent to which users explicitly make use of the choice provided by GDPR cookie notices* and choose to block third-party tracking. I examine this using the anonymised cookie data from one year of browser histories of the UK users in this study. Fig. 7.6 shows that although there was a brief reduction in the number of third-party cookies when GDPR was introduced in May 2018, the overall number of cookies among the 9 UK users has stayed relatively the same between Jan 2018 and Jan 2019. The reductions between Mar 2018 and Jun 2018 appear to coincide with the beginning of the preparations for GDPR cookie compliance and the cookie consent manager rollouts of the widely used OneTrust [One18] (Mar 2018) and TrustArc [Tru18] (Apr 2018) for GDPR compliance, and similar reductions also reported by others[LGN18]. However, Do Not Track cookies and GDPR consent cookies expire; cookie caches get cleaned etc, and *it appears that users in this study have subsequently mostly chosen default settings or have made choices that do not increase their privacy – there is little change in the numbers of third-party cookies per website visited between early 2018 and early 2019.* Table 7.3 shows how the numbers of cookies varied for selected sites of different `Alexa.com` ranks between Feb 2018 and Feb 2019, with a minimum being seen around the time GDPR introduced in May 2018. Interestingly, users in China experience *fewer* third party cookies throughout the duration.

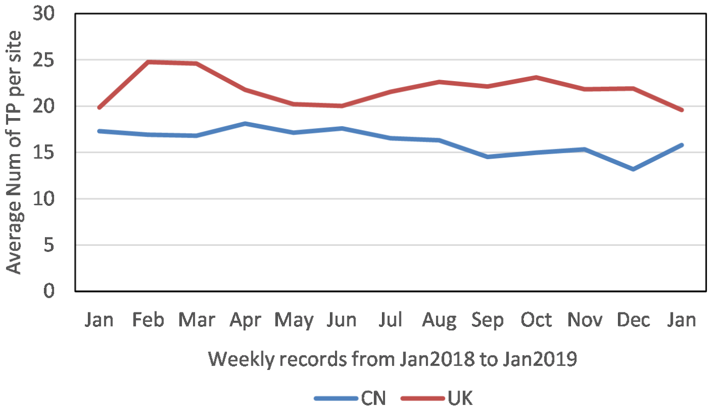


Figure 7.6: Average number of third parties per site, based on weekly browsing records of UK and China participants.

	Site A (top100)	Site B (top200)	Site C (top300)	Site D (top400)	Site E (top500)
Feb., 2018	13	14	20	21	37
May, 2018	8	8	16	17	29
Feb 2019	12	8	22	18	32

Table 7.3: Number of cookies on websites visited by real users.

7.5 Discussion and Conclusions

In this chapter, I took an in-depth look at the effect of GDPR, which requires cookie notices when sites are using third-party cookies that collect personal data. I find that although UK-based websites comply in general (i.e., serve some form of cookie notice), non-EU sites are less likely to offer fine-grained choices for users to decide their privacy preferences. Availability of choice also varies across different categories of websites, with adult websites being the least likely to offer a cookie notice, but also with many fewer third-party cookies than other categories such as news websites.

Fine grained choices are not necessarily what is "best" for the users: First, though UK websites are meeting the cookie consent requirement by presenting users with a choice, this choice can be a false one – if default choices are accepted, it could sometimes lead to higher numbers of third-party cookies than before. Second, by studying the numbers of third party cookies in real users' browsing histories, I find that GDPR has had little long term effect on the numbers of cookies. **The finding**

might be limited because it is drawn based on a small user base. However, it could be used as the basis of the future general conclusions. In practice, the choices, when offered, can be very fine grained (e.g., Fig. 7.1 (b)), allowing users to opt out of cookies from specific third parties that are being used by the website while still allowing them to opt in for cookies from other third parties. I speculate that users may be fatigued by the effort of having to choose their privacy preferences on every website they visit, and end up accepting the default choices offered by the websites (which in a majority of sites, is to have tracking turned on). Interestingly, users in the UK appear to have *larger* numbers of third party cookies than countries like China. Unfortunately, tracking is the default on many sites where users are not given a choice at all, and the only real choice for users appears to be a forced one of either accepting tracking and third party cookies, or not using the website at all.

In summary, I find that by and large, the relationship between website operators and users remains unbalanced, and GDPR may in practice be falling short of the level of protection that it aims to deliver.

Chapter 8

CookieCutter: Privacy Gain Without The Pain Of Manual Cookie Consent Management

It is time to stop the anarchy on the Interne.

—— Alexander Lukashenko

Summary

In Recent times Data privacy protection laws like the General Data Protection Regulation (GDPR) attempted to regulate the flow of information in the digital realm. Within its jurisdiction, GDPR introduces the cookie consent notice banner in websites, where the websites needed to ask *informed consent* from the users before they can set cookies in users' browsers and track them. Under the hood, these banners leverage GDPR consent cookies, a set of cookies that stores the user's cookie setting preference (e.g., whether to allow advertising cookies) for a particular website. However, current approaches to cookie consent notifications often fail to protect users due to reasons like consent fatigue and the lack of consent withdrawal mechanisms.

To address this problem, I motivated, designed, built and deployed CookieCutter, a novel usable browser extension for assisting users with setting the current values of GDPR consent cookies. CookieCutter uses a machine learning model to identify GDPR consent cookies with 92% F1-score. Furthermore, as a first tool CookieCutter also automatically sets the correct values of GDPR consent cookies (using a manual curation as well as *peer sourcing* approach). In effect CookieCutter minimised privacy-violating cookie-based tracking as desired by GDPR. I developed CookieCutter with 13 participatory design interviews, which revealed key design principles and a need for automated assistance for GDPR consent cookies management. The deployment of CookieCutter in the wild resulted in obtaining 96 users from 25 countries. The anonymised data collected from these users as well as their feedback (via optional online surveys) demonstrated that CookieCutter automatically assist users to protect user privacy effectively in more than 85% of the visited websites which use GDPR consent notices.

8.1 Introduction

Introduced in May 2018, the General Data Protection Regulation (GDPR) has been seen as a resounding success whose template has started to be copied in many other legal jurisdictions around the world such as Brazil (for LGPD), South Africa (for POPIA), etc. One of the requirements of GDPR is that any establishment that wishes to collect data about a so-called "natural person" needs to first obtain their consent. This person, also called the "data subject", has a number of other associated rights including the right to withdraw a consent previously given, and the right to ask for erasure of the data collected about them and the rectification of any errors therein.

GDPR has had far-reaching consequences in many aspects of the lives of people living within its jurisdiction[1]. The most visible consequence (to ordinary citizens) are

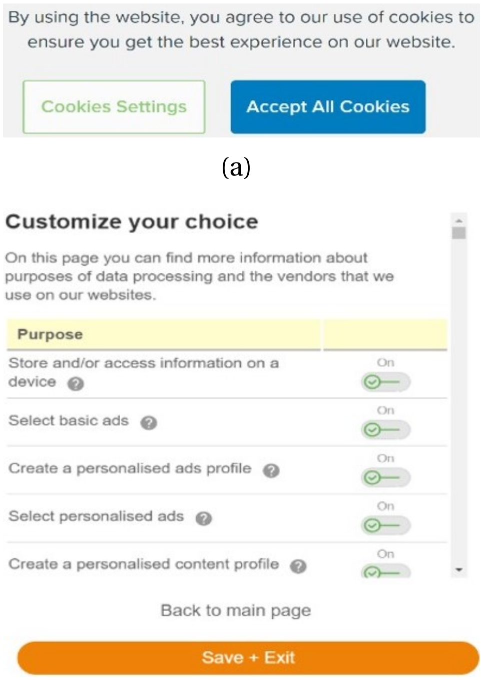
[1] GDPR's jurisdiction is the EEA, which consists of the EU and a few other countries such as Norway. The UK has also adopted GDPR as the "UK GDPR" after Brexit, although this may be reviewed after the end of the transition period.

the cookie consent banners have become ubiquitous within the jurisdiction of GDPR (EU and a few other countries such as UK): As a consequence of the GDPR and the older ePrivacy directive[2], website operators are now required to ask for user consent before collecting data about them, *e.g.,* for advertising and tracking, or even for analytics to improve website performance. Although these so-called *cookie consent banners* (or *notices*) have led to several fundamental changes in the web tracking industry, several studies point out that users struggle to express their privacy requirements and user tracking is still rampant[SK19, SM19a, SRDK$^+$19a, HS19, DMU$^+$19, Kri, Xue20, PPKM21, UTD$^+$20].

Scholars have started to attribute this return to tracking status quo to the notion of "consent fatigue" [HH19]: Tired of having to carefully navigate the cookie consent banners to set the most private option, even privacy conscious users simply choose the default option which "accepts all" trackers [KS21]. User fatigue caused by the strain of having to constantly enforce their privacy rights is known to have a stronger effect on privacy behaviours than the privacy concerns and stances that users may have a priori [CPJ18]. Websites also commonly employ "dark patterns" [NLV$^+$20] that make it difficult for users to choose the most beneficial option for them. Together, consent fatigue on the part of users and the adoption of dark patterns by websites completely defeat the concept of "meaningful consent": despite the strong privacy protection afforded in theory by GDPR, tracking has come back to pre-GDPR levels [DUL$^+$18, HS19, DMU$^+$19, SM19a].

This current situation is exacerbated by the way cookie consents work: When a user first visits, websites present users with a banner that seeks their consent to collect data about them (Figure 8.1). However, once they give this consent (even if that "consent" is simply a click on the "Accept All Cookies" button merely to make the banner go away), the consent is recorded and the banner which seeks the consent is not shown to the user on subsequent visits to the same website. Although this decision to not repeatedly seek a user's consent is critical to ensure a good user experience, this also

[2] https://gdpr.eu/cookies/



By using the website, you agree to our use of cookies to ensure you get the best experience on our website.

Cookies Settings | Accept All Cookies

(a)

Customize your choice

On this page you can find more information about purposes of data processing and the vendors that we use on our websites.

Purpose	
Store and/or access information on a device ?	On
Select basic ads ?	On
Create a personalised ads profile ?	On
Select personalised ads ?	On
Create a personalised content profile ?	On

Back to main page

Save + Exit

(b)

Figure 8.1: Cookie consent notices presented to the users in GDPR-compliant websites.

means that on most websites, it is extremely hard for a user to change their mind once they have provided consent to be tracked. Note that this is in direct contravention to the requirements of GDPR, which explicitly provides data subjects with the right to withdraw their consent. Again, scholars have identified the difficulty of withdrawing consent as an important issue [KS21, LCJ^{+}20].

To address these problems, I have designed, built and deployed a browser plugin called CookieCutter that is publicly available on the web extension stores of Google Chrome, Mozilla Firefox and Microsoft Edge. CookieCutter is currently being used in-the-wild by 96 people from 25 countries. Survey responses from these users indicate that nearly all (over 95%) strongly agree or agree that CookieCutter allows them to effectively control their data. Importantly, *a significant number (31) of users of CookieCutter are from 7 countries outside the jurisdiction of GDPR. CookieCutter provides these users with similar protection as for 65 users in the EU*. This is because many websites deploy the consent management scripts even outside the EU, to meet the requirements of GDPR (Details in §1.2, Problem 3, ***Cookie consents globally.***). Thus, simply by setting the right values for a handful (typically 1–3) of cookies that manage GDPR consents, I enable users across the world to enjoy similar protection from tracking as users within the EU.

CookieCutter uses a mixture of carefully handcrafted regular expressions and machine learning **instead of exacted values** to automatically identify the names of GDPR consent cookies (F1-score = 0.89)**, which could provide more flexibility and control while performing pattern matching**. Just knowing which of the several (tens and sometimes over 100) cookies on a website correspond to GDPR consents will allow users to selectively delete those specific cookies and force the website to present its cookie consent banner again and renegotiate cookie consents with the user. This addresses the problem of consent withdrawal.

CookieCutter goes beyond automatic withdrawal of previously given cookie consents and automatically *sets* the values for GDPR consent cookies that maximise users' privacy. This is achieved through a combination of different manual but scalable solutions. Since each website can take its own approach to managing consent, there is no uniform or systematic way of setting the right values for GDPR consent cookies. The solution to this problem takes two complementary approaches: First, I observe two forms of consolidation in the way cookie values are set – (*i*) the emergence of Consent Management Platforms (CMPs), which are third party libraries that help manage user consents [SNT+21]. All websites using a particular CMP use similar GDPR consent cookies ˙(*ii*)More popular websites are visited significantly more often by users. I take advantage of these two forms of consolidation *handcrafting* the values to set for these patterns of usage: I handcraft the values for the top 25 CMPs observed among popular websites. I then find that among the Alexa UK Top500 websites, 90 websites do not use any CMP and instead deploy a custom solution. I handcraft the values to set for these highly visited websites. Together these 115 (90+25) handcrafted values for GDPR consent cookies collectively account for over 62% of the websites visited by the in-the-wild user base.

The second approach complements the handcrafted values by allowing *allowing users in-the-wild to contribute values*: When the browser extension detects GDPR consent cookies but does not know what values to set for those cookies to maximise privacy, it allows users to click on the cookie consent banner to manually set the values. It then asks the user whether they would like to contribute those values to help other users, thus creating new values for GDPR consent cookies which others can through *peer-sourcing*. Note that only GDPR consent cookies (rather than all the cookies set by the site) are sent back to the central database. Users are also offered the option to manually edit the list of values sent back, giving them full control to ensure that no sensitive data is shared. Multiple users may contribute values back to the database, and some of these users may have not chosen the most private or "reject all cookies" option. To mitigate this problem, the central database records the decline in the number of cookies from each contributed set of cookie values and distributes to other users the cookie values which afford the largest decline.

The rest of this chapter is structured as follows: Section II discusses related work. Section 8.2 discusses the problems with current cookie consent notices and presents a solution sketch for how cookie consents can be managed automatically. Section 8.3 then takes a participatory design approach to understand user perceptions about automatic management of consents on their behalf and derives design principles to guide the development of CookieCutter. Section 8.4 presents the design of the extension CookieCutter. Section 8.5 compares against other such extensions and with other means of minimising tracking, such as ad blockers. Section 8.6 evaluates the deployment of CookieCutter in-the-wild and Section 8.7 concludes.

8.2 Automatic management of GDPR Consent

8.2.1 An alternative to cookie consent notices

Cookie consent notices work by actually setting a small set of additional cookies whose values record users' preferences on what data they permit the websites to collect, whether for tracking or other purposes. I ask a simple and straightforward question: If I am able to *programmatically* set the right values, does this prevent tracking?

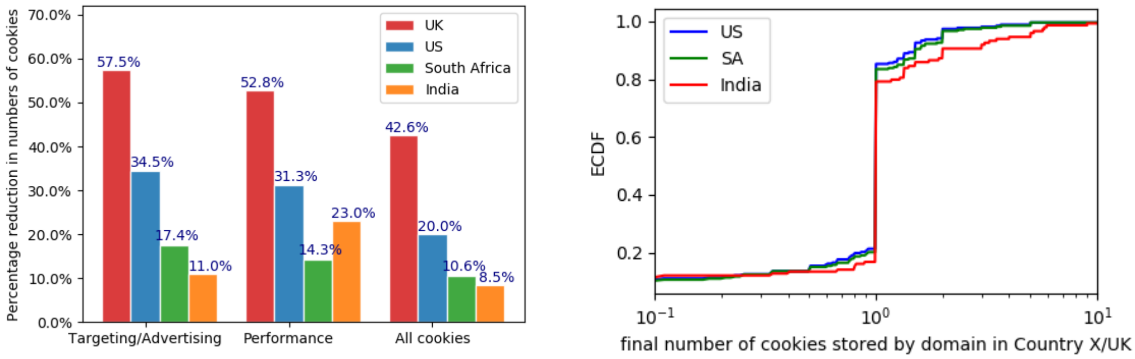
To test this, I exhaustively examined all the Alexa top 100 websites in the UK (rankings as of Jan., 2021). Interestingly, only 55 these websites present users with cookie consent banners, although these sites collectively use the top 8 CMPs, representing over *65%* of the CMP market. Determining whether the other 45 may be in violation of GDPR or may not be collecting Personally Identifiable Information (PII) is outside the scope of the study[3]. Instead, I focus on the 55 websites which do ask users for consent to collect their data and manually record the names and values of the cookies which are set when the most private CMP options are chosen by the user. I call these cookies as *GDPR Consent Cookies.*

I then create a new browser profile and *pre-populate* it with the GDPR Consent Cookies previously identified. Next, I visit the 55 websites again and confirm that the user is not presented with a CMP banner in any of the websites. I compare the numbers of (non-GDPR Consent) Cookies set by these websites (for tracking, analytics, etc.) when a user manually chooses the most private option by clicking on the cookie consent banner vs. the number of non-GDPR Consent Cookies set when the GDPR Consent Cookies of these websites are pre-populated for the user. I find that both methods yield similar numbers of cookies (Table 8.1 shows a median difference of 0 for all categories of cookies and a maximum difference of 3 analytics-related cookies), which confirms that *programmatically setting cookies not only saves users from having*

[3] although I note that 55 is a slight improvement over a previous 2019 study that reported only 42 of the Alexa UK Top100 sites offered users a cookie consent notice with choices [HS19]

Cookie type	max	mean	median	min
Advertising	1	0.2	0	0
Analytics	3	0.6	0	0

Table 8.1: Difference in numbers of advertising and analytics cookies when a user manually chooses options via a cookie consent notice vs. when GDPR consent cookies and their values are pre-populated in their browser's cookie jars.



(a) Percentage reduction in cookie numbers in four countries after applying the proposed approach

(b) CDF of number of cookies left after applying the proposed approach, for different websites accessed from country X, shown as a fraction of the number of cookies for the same website accessed from UK

Figure 8.2: Cookies on 55 Alexa Top websites when accessed from UK, USA, India and South Africa.

to choose the right GDPR consent options on each website but also achieves similar effects in decreasing data collected about the user (solving problem 1).

Next, I check whether the technique of pre-populating GDPR consent cookies in users' browsers works *outside EU locations*. Article 3 of GDPR stipulates that websites operated from or controlled by EU/EEA-based establishments should obtain consent from all users globally. Similarly, Article 3 also states that establishments outside of the EU must respect GDPR for data subjects in the EU for "monitoring of their behaviour as far as their behaviour takes place within the Union", thus I would expect (and do observe) that global websites to also deploy cookie consent functionality from a vantage point within the GDPR jurisdiction region.

To quantify this systematically, I use a VPN solution with four different end points — UK, USA, India and South Africa (locations chosen to be populous countries representative of the different continents of the world; but where reliable exit points were available for the chosen VPN solution. For Asia, India was chosen instead of China because many websites are not accessible behind the Great Firewall of China. The VPN did not support exit points in Australia or South America). I then visit the same 55 Alexa top sites as above, first without setting any GDPR Consent cookies and then after setting GDPR consent cookies. Figure 8.2(a) shows the percentage reduction in cookies in each country, showing that *the approach can have an effect globally, even outside of the EU.* Although it appears that some countries have a greater amount of reduction, this is simply because of the larger number of trackers initially before the protection is applied. This is corroborated by previous studies for example [HdTS20] which shows that UK has more trackers than the US, which means that the reduction in the UK after removing all trackers would be proportionally more than in the USA. Similarly [SM19b] shows that UK has more trackers than the USA which in turn has more trackers than South Africa, followed finally by India. This rank order corresponds to the different percentage reductions I see in Figure 8.2(a). Figure 8.2(b) confirms that after applying the approach, most websites have the same number of (essential) cookies left in the three countries outside the jurisdiction of GDPR (*i.e.,* USA, India and South Africa) as from UK vantage points (which is in the jursidiction of GDPR). This suggests that simply setting GDPR consent cookies could afford users in countries around the world similar privacy protection as provided by GDPR in the EU (Problem 3).

Finally, when I manually remove the identified GDPR consent cookies and refresh the web page, the GDPR consent banner pops up again to offers a way for users to revisit or change their previously set cookie consents. Given all of the above evidence, it indicates that the GDPR consent cookies that governs the ability to assign cookies to websites is indeed captured.

8.3 Acceptability of Automatic Consent Management Solutions

In the last section I manually identified the cookies used by a website for recording GDPR consent. The investigation demonstrated that simply copying the values of these cookies in a new browser profile achieves the same results as the user setting these cookies through a cookie consent banner. Thus, I reason, if the detecting cookies and setting their value could be automated, I can manage to maximise users' privacy without causing consent fatigue; moreover this solution might apply across the globe, even outside the EU.

8.3.1 Towards automatic cookie consents: a solution sketch

Recent prior work uses machine learning to classify cookies [HSM21] with high accuracy. Thus, one can imagine an AI/ML model that can be trained to identify the cookie names corresponding to GDPR consents. The manual examination from the 55 websites suggests that the values of GDPR consents are fairly formulaic (e.g., the value of consent cookie for OneTrust (*OptanonConsent*) composed of the following required fields: "isIABGlobal(Boolean)"+"datastamp"+"version of embeded OneTrust"+"path of current page"+"consents to each cookie category(0/1)"+"Consent Id"); thus for any given GDPR consent cookie, the values set by one user may be copied by other users, essentially *peer-sourcing* the right values for GDPR consent cookies to avoid the user having to set these through the CMP used by the website.

Yet, deploying such a solution may completely defeat the purpose of *informed* consent — would users accept an AI/ML solution that sets GDPR consent cookies on their behalf, with an aim to maximise their privacy? Would users be willing to accept the cookie values of other users? In the rest of this section, I test these questions with a small user study. In the user study along with a survey, I use participatory design principles (via detailed interviews with prototypes) to develop and refine the design of a browser extension to automatically set GDPR consent cookies [SWS15].

8.3.2 Study design

In this study I show the participants a preliminary version of a browser extension popup and conducted an online interview where the participants wrote and send me answers to specific questions (while online). First, I asked the participants to install a version of the extension. Then I focus on key components of the extension (two main panels visible to the users and an advanced toolbar) and asked them specific questions regarding the desired functionality (including their perception regarding usage of Machine learning and peer sourcing) as well as interface elements of this extension (questions from the study are in Appendix A.4). I synthesised key themes from user feedback using a thematic analysis [BCHT19] and incorporated the feedback in the extension within four weeks. Then I again contacted the participants and verified from the participants if the final design incorporated their feedback.

I recruited a total of 13 participants for this study using student and associate mailing lists available at the lead author's university (addressing the study as an exploration into improving their web browsing experiences via an extension). I also asked the mail-recipients to check if their friends are interested in this study. Ultimately 13 participants self-contacted me via email. All of the participants finished college, were fluent in English and between 20 to 35 years of age. The participants were originally from four countries spanning both Asia and Europe. There were 9 male participants and 4 female participants. Each study in total took 30 minutes. I did not provide any monetary compensation to the participants. The study is approved by the Institutional Review Board (IRB) of the lead author's institution. I did not keep the email id or any other identifier associated with user feedback, protecting the anonymity of the participants.

Table 8.2: Participant responses on acceptability of automated methods for managing cookie consent in the interview.

Queries	Responses (13 users)		
Comfortable with AI/ML methods to set privacy	YES (13)	NO (0)	
Comfortable using other users' privacy settings (peer sourcing) to set privacy	YES (8)	NO (5)	
Preference of AI/ML and peer sourcing	AI/ML (9)	peer sourcing (4)	
Which cookies would you prefer an automated mechanism to block on your behalf?	ad cookies (5)	ad + analytic (1)	all except essential (7)

8.3.3 Acceptability of automatic cookie consent management

I qualitatively analysed the responses of 13 participants. **Some of participants further state their perceptions on GDPR, helping us deepen the understanding of user requirements with a high level of credibility. (I confirm the ethical commitment to represent the data collection during usability research in an objective and anonymous manner.)**

> ***I like visual confirmation that every possible cookie has been blocked. without reject all is far too tedious, the binary choice leaves me wondering if I have been allowed to block performance and functional cookies, cookie categories is nice, but I still prefer to see the list or all providers with opt out all, plus I think it's important for people to see the list of providers so they can see just how ridiculously large and unnecessarily complex the business of data farming has become. The slider option is poorly designed and doesn't allow for the selection of non-essential 1st party cookies.***
>
> **——— By participant 3**

According to the quotation above, it could be found that "Full vendor list" sometimes brings users a greater sense of security and control. Even though "binary

option" offers a shortcut option for one-click rejection, users still believe in a detailed vendor list. Implicit cookie banners would increase the distrust between the user and the website.

> ***I don't like the full vendor list its too long, full of clutter and not easy to use.***
>
> **——— By participant 7**

But some participants, such as participant 7, are dissatisfied with the time-wasting vendor list. As well as ignoring the regular cookies cleanup, this cohort of users also normally has no previous experience with other privacy-oriented related ad blockers/browsers/search engines. Thus, they tend to rarely check the details when consenting to the GDPR consent banners.

To gain insight into user needs for cookie banners (management), I took an thematic analysis approach [BCHT19] to extract broad themes regarding two high-level questions—(1) are users comfortable with automated methods to help with cookie banners? (2) How can I change the functionality and/or interface to make the final extension usable. To answer the first question, one author first created four questions to guide the coding process (Table 8.2). Then I used focused coding [Sal15] to uncover relative presence of themes regarding usage of automated techniques. Table 8.2 present the results.

Comfort with automated methods to assist with cookie consent: All the participants were comfortable with automatically setting cookies for them with a preference towards using AI/ML methods, however, four out of thirteen participants preferred peer sourcing. Investigating more, I found that the slight suspicion of the minority (5 users) towards peer sourcing was rooted in a distrust on the cookie settings provided by non-GDPR CMP providers and on the effectiveness of peer sourcing.

Desire to have "Reject all" cookies as default: majority of participants (seven out of thirteen) want to disable all cookies unless necessary and all of them wanted to disable advertisement related cookies. However, some participants also wanted to allow non-ad cookies. On probing further, I realised these participants believed that analytical and functional cookies do not cause great harm to privacy and hence they desired to allow those cookies. However, their perception is not correct—some third-party websites (such as web analytics companies) can aggregate the data collected from multiple first-party websites using their services, track users across websites and in effect harm their privacy. Even, functional cookies can include social sharing features, in effect making social networks like Facebook as tremendously powerful data aggregators.

Summary: Consequently, from the qualitative analysis, I decided that participants are often comfortable with automated assistance with cookie consent banners to enforce GDPR protection. Moreover, I decided to adapt the "reject all" setting as the default option for the extension. Next, I checked the themes regarding the guiding design principles to make the app usable.

8.3.4 Suggestions for UI design

Next, to answer the second question regarding making the final browser extension usable, I perform another thematic analysis. One author specifically identified the quotes from the participants where they suggested to "improve" or "change" the interface. Then, two authors collaboratively merged similar changes from the quotes and identified four key design principles for the extension.

Design principle 1: Keeping only key information in main interface: The main panel shown to the users presented all the details about the cookies stored in the current website and the browsing history. However, it caused information overload. Thus I divide this detailed information into a second-level page (one click away from main panel of the extension pop-up) to distinguish professional and normal users' requirements for functions of the browser extension.

Design principle 2: Simplified automated privacy protection by default: In the first iteration of the plugin (used in participatory design) I used double-slider options (see Appendix A.1) to enable automatic protection application for each website individually (default was disabled for maximum user control). However, participants desired that the extension interface should support automatic privacy protection *by default* rather than manually having to enable automatic cookie consent setting for each website. I incorporated the advice and set the sliders accordingly.

Design principle 3: Clearing previous GDPR consent cookies: The participants identified that for first time installation of the extension, they desire to have an option to remove the previously consented cookies (before they start using the extension). Otherwise, they argued that, there might be first-party websites that have been accessed before the installation which can track them, resulting in being excluded from automatic protection of the extension. I have added a button that provides users with the option to delete all previously set GDPR consent cookies.

Design principle 4: Privacy worries in peer sourcing mechanism: The proposed solution has peer sourcing component in which users help each other by contributing the values they set manually for GDPR consent cookies on a given website. Some participants were worried whether these contributed values for cookies might contain sensitive data about them.

To address this issue, I decided to provide users details of the contributed cookie sets before they share those cookies. In fact, before making a contribution of the custom cookie setting through the browser extension, I provide details of the cookie object shared (based on which the users can cancel or not share one or more cookies that may be relevant to GDPR consents). (Figure A.2 in Appendix A.2).

Armed with these design principles, as well as being motivated by user need for automatic cookie consent management, next I detail the design of the final browser extension.

8.4 Design and Development of CookieCutter

As outlined in §8.3.1, automatic management of GDPR consent cookies requires being able to identify the names of these cookies (§8.4.1) and setting the right values that maximise privacy (§8.4.2). I have implemented these functionalities (incorporating the feedback from the participatory design exercise of the previous section) in a browser extension called CookieCutter that is available on Google Chrome, Mozilla Firefox and Microsoft Edge extension stores.

8.4.1 Detecting GDPR consent cookies

To understand how to recognise which cookies are GDPR consent cookies, I extend the manual exploration in §8.2.1 to the top 500 websites, systematically deleting cookies to identify the subset of cookies that are set when a user selects by hand the options allowed in the website's cookie consent notice. I confirm that these are indeed the set of cookies that record a user's consent to be tracked by deleting these cookies and ensuring that the cookie consent notice pops up again.

The study reveals some interesting patterns: 277 (*i.e.*, over 50%) of the UK top500 websites do not present the user with any consent notice whatsoever. This represents a slight fall from the 45% of the top 100 websites without cookie consents (§8.2.1). Of the 277 sites which do offer cookie consent choices, 93 appear to have a custom or "home grown" solution whereas the remaining 130 sites use one of 25 different Consent Management Platforms (CMPs) such as OneTrust, Quantcast and Didomi. A similar trend is seen in the top 100 sites as well: among the 55 which have cookie notices, there are 17 custom solutions but a much larger majority (38 sites) use one of 8 different CMPs.

CMPs are third party libraries that can be easily integrated into websites to manage cookie consents. The library nature of CMPs means that that *multiple websites tend to use the same cookie names or follow a similar template* in managing GDPR consents. *e.g.*, sites using Quantcast/IAB use "_cmpRepromptOptions", OneTrust uses "Optanon" as a prefix to several cookie names (like "OptanonConsent", "OptanonAlertBoxClosed"), etc. Furthermore, names of different CMPs (e.g., Evidon, Didomi, Axeptio) are also frequently used in the cookie names. I exploit these common patterns by capturing them as a regular expression (which I term as the *GDPR* *regex*. See Appendix A.7 for details). This pattern recognises the GDPR consent cookies of all the 25 CMPs used by Alexa top500 sites, plus a few other patterns (*e.g.,* "euconsent" or "hasseencookiedisclosure") that could be deduced as being related to GDPR consent management. **TThere also exists generic cookie names (e.g., settings, preferences) that websites do not usually use for different purposes because they make use of the common APIs of the specific CMPs.**

To detect additional GDPR consent cookies which are not directly captured by the *GDPR regex*, I adopt a Machine Learning (ML)-based approach and build a binary classifier that labels cookies as GDPR consent cookies or not, using the manual labels from the Alexa top500 sites as ground truth. The main challenge in applying standard ML techniques is that in any website, *most* cookies except around 1–3 are not GDPR consent cookies. This creates two problems: (*i*) a large class imbalance, which can be addressed by creating a balanced dataset with equal numbers of GDPR consent cookies and other cookies and (*ii*) a lack of sufficient numbers of positive labels even in the balanced dataset. To address this, I programmatically visit every website in the Alexa Top20K using Selenium, and use the *GDPR regex* to identify additional GDPR consent cookies. I then create a balanced dataset from all the GDPR consent cookies and other cookies identified from all the 20K websites visited.

I noted via manual inspection of the collected cookies that some parts of cookie names can be meaningful for both positive and negative labels. Thus I decided to use features extracted from these names for categorisation. First, I removed all numbers from each cookie name ((e.g., ADS_324 became ADS_). Next, I tokenise these names using punctuation characters (e.g., %, ~, ., _, -). Thus, at the end of preprocessing and tokenisation, a cookie with the name *gdpr-track-status45* will be split into tokens *"gdpr", "track", "status"*. Furthermore, I split the resultant token using capitalisation (i.e., AnalysisUserId → [Analysis, User, Id]) . Finally, I case-folded the resulting tokens.

I then explore five kinds of classifiers to process the binary classification based on the features extracted from cookie names. Specifically, I evaluated Softmax Regression (Multi-layer perceptron or MLP), Support Vector Machine with a linear kernel (SVM),

Table 8.3: Recall, Precision and F1-score of for different classification models to GDPR consent cookies.

Algorithm	Precision	Recall	F1-score
SVM (linear kernel)	**0.927**	**0.932**	**0.930**
Multi-layer Perceptron (MLP)	**0.927**	0.923	0.925
Random Forest	0.923	0.917	0.920
Bernoulli Naive Bayes (BNB)	0.920	0.917	0.918
K-Nearest Neighbors (KNN)	0.918	0.912	0.915

K-Nearest Neighbors (KNN), Random Forest, and Bernoulli Naive Bayes (BNB). I used a 5-fold cross validation with 80-20 split between training and testing data. I used overall (Micro) precision, recall and F1-score over all-classes to report the accuracy of categorisation for all of the models in Table 8.3. All methods achieve an F1-score of more than 0.9 but I choose to bundle this pre-trained SVM model in the extension as it achieves a slightly higher recall and F1-score compared to other algorithms.

The final solution consists of using *GDPR regex* to identify well known GDPR consent cookies, with the above pre-trained SVM model as backup to catch a further set of cookies. The solution is expected to perform well in the limited setting of the curated dataset, as (*i*) the *GDPR regex* was constructed using the dataset and (*ii*) the SVM model has been trained using the dataset as ground truth. I next check whether this approach might generalise to other websites not seen before. I randomly select $n = 60$ websites, choosing 10 each from the Alexa 1-100, 101-500, 500-1000, 1K-10K, 10K-100K and 100K-1Million ranks. Table 8.4 shows that the combined *regex*+ML approach performs well, with an average F1-score of 0.89. The F1-score suffers in the 500-1K range of websites and the recall is lower in the 100K-1M range because of three less common CMPs Webedia, Axeptio and RODO, which are present in the websites chosen in these rank ranges and which have not been seen before. The browser extension now incorporates an updated version of the *GDPR regex* that can identify these CMPs' cookies as well. I also allow users to report websites on which there is a cookie consent management solution but which the extension does not recognise. Using such user feedback, I can improve the extension in future versions.

Table 8.4: Recall/Precision rate and F1-score for GDPR consent cookie recognition across `Alexa top-1M` websites.

10 sites/topK	Recall	Precision	F1-score	CMPs (in random 10 sites)	Detected by Regex
top1-100	1	0.88	0.94	Evidon, Quantcast, Didomi	66.67%
top100-500	0.94	0.76	0.84	OneTrust, Evidon, Quantcast,Admiral	85.71%
top500-1k	0.83	0.85	0.8	OneTrust, Evidon, Webedia	77.78%
top1k-10k	0.83	1	0.91	OneTrust, Quantcast	75%
top10k-100k	0.89	0.9	0.89	Quantcast, Cookiebot, ConsentManager	60%
top100k-1M	0.8	1	0.89	Axeptio, RODO, Quantcast	50%
All	0.90	0.87	0.89		69.19%

As shown later (Fig. 8.5), users have found that around 4.83% of the websites they have all seen collectively have consent notices which the extension does not recognise.

8.4.2 Setting GDPR consent cookie values to maximise privacy

Identifying which cookies are GDPR consent cookies is only part of the problem. Once these cookies are identified, I also need to set the "right" value that prevents user data from being collected.

The first is based on the key observation that the complexity of consent management has led to a consolidation of functionality and the birth of a new industry of Consent Management Platforms (CMPs) [SNT+21]. CMPs are third party libraries that can be deployed by websites to manage their cookie consents. Popular CMPs are used by several websites and each CMP has a well-known set of GDPR consent cookies. Thus, by handcrafting the right cookie values for the top 25 CMPs, CookieCutter is able to handle a large number (62.32%, *cf.* Fig. 8.5 for details) of all the websites visited by the user base (currently nearly 100-strong).

The second solution is to handcraft the correct GDPR consent cookies values for highly popular websites which are visited by many users. In the case, I look at the 500 most popular websites and find that 223 of them use one of the 25 CMPs discussed above. Of the remaining, 197 do not use cookie consent banners, either because they do not collect personal data or perhaps in violation of the requirements of GDPR. In either case, tracking and data collection (if any) by these 197 websites cannot be manipulated through setting GDPR consents appropriately. The remaining 90 websites use their own custom cookie consent notice rather than a widely used CMP. I handcraft the right values for their GDPR consent cookies and find that this accounts for a further 11.11% (*cf.* Fig. 8.5) of users' browsing history.

The third and final solution is to employ *peer-sourcing* and allow users to help each other: If CookieCutter detects some GDPR consent cookies but does not know what values to set for them, it waits for the user to set these values manually through the cookie consent banner and then asks the user whether they would like to contribute these values to a central database. All other users of the extension can then benefit from this by simply copying the values contributed by the first user To avoid sharing sensitive personal data, the contributing user is shown the values which will be shared and the user can optionally delete or share them.

In a minority of cases, CookieCutter does not detect the names of the GDPR consent cookies. This may be because the website does not employ cookie consents, or because it only provides a notice to users that their data is being collected, without offering the choice of refusal (although this might violate the rights afforded by GDPR to a data subject) or because the website does deploy a cookie consent notice but its GDPR consent cookies is not detected by mechanisms. I allow the user to notify me if they can see that a website has a cookie notice but is not being detected by me. Collectively this corresponds to about 4.83% (*cf.* Fig. 8.5) of websites visited by the in-the-wild user base and I am using this to improve the extension.

8.5 Extension evaluation

I next compare CookieCutter with two other solutions with similar intentions. The first §8.5.1 is an extension called Minimal Consent, which currently seems to be the state of the art for automatically setting cookie consents. The second (§8.5.2) is ad blockers, which do decrease tracking as a side effect of blocking trackers.

8.5.1 Comparison with baseline

Extensions such as MinimalConsent [Mad20], Consent-o-matic [Uni19] by Aarhus University, PrivacyCloud Consent Manager [Pri20] and I Don't Care About Cookies [Kla21] provides similar functionalities as CookieCutter. However, the extension has an advantage in that it uses machine learning and regular expressions, apart from handcrafted expressions which are commonly used in these extensions. In addition, I believe the list of handcrafted expressions is more extensive than in other implementations.

In order to compare existing tools and the extension, I install Minimal Consent as an example and then visit UK Top500 and the Global Top500 websites. Table 8.5 shows the relative numbers of each of the Global and UK Top500 categories that do not support GDPR or offer a cookie consent banner with "no choice" (*e.g.*, "if you continue using the website, you accept to being tracked"). These extensions (and ours) are only able to work on websites that do offer users a choice in how they are tracked. Ultimately 223 websites from the UK Top 500 and 91 websites fromthe Global Top500 websites offer users a GDPR consent notice with choice and can benefit from extensions that automatically manage cookie consents. The table also shows that many of these websites use local storage apart from cookie storage to record consents. I believe only the extension is able to clear previously set cookie consents and furthermore it does so not only from cookie stores but also from local storage.

Table 8.5: There are 91/223 websites in Global/UK top500 provide both GDPR notice and explicit GDPR options, but 11 websites can not be accessed from UK.

Datasets	no GDPR	GDPR no choice	GDPR with choice		no access
			cookie storage	LocalStorage	
GlobalTop500	248	144	82/91	15/91	17
UKTop500	95	182	205/223	55/223	/

Among the 223 UK websites and 91 global websites that provide users with GDPR options, Table 8.6 (respectively Table 8.7) display the number of websites from the UK Top 500 (respectively Global Top500) that can be automatically protected with the minimal privacy from different CMPs and from websites that do not use a CMP but use a custom solution. CookieCutter is able to support over 98% of the websites that support user choice in GDPR consent management whereas Minimal Consent works only for 41.25% of UK top500 sites and only 34.07% of the Global Top500 sites.

Table 8.6: Comparison of supported and unsupported GDPR consent management among the 223 **UK** websites. (s|f) refers to the number of websites that are supported/failed to support. **(CC: CookieCutter, MC: Minimal Consent, COM: Consent-O-Matic, CM: Consent Manager, IDC: I Don't Care About Cookies)**

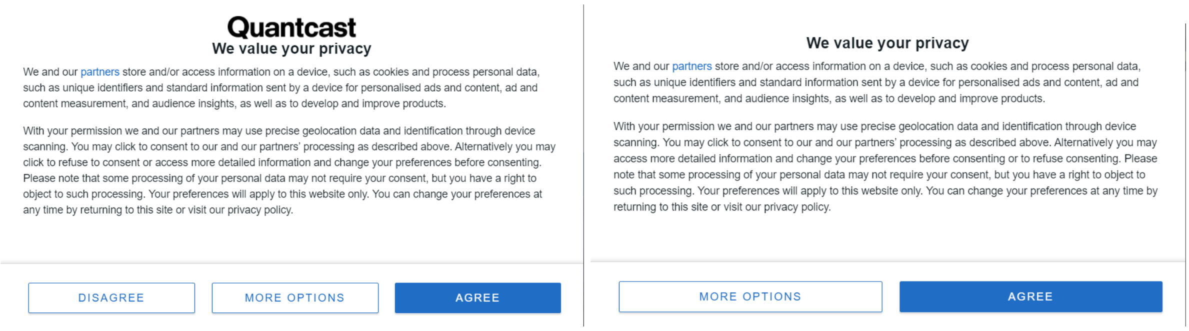
CMP	CC (s|f)	MC (s|f)	COM (s|f)	CM (s|f)	IDC (s|f)
ConsentManager	1 | 0	1 | 0	0 | 1	0 | 1	0 | 1
Vivendi	1 | 0	0 | 1	0 | 1	0 | 1	0 | 1
CookieBot	2 | 0	2 | 0	1 | 1	2 | 0	1 | 1
Rodo	2 | 0	0 | 2	0 | 2	1 | 1	1 | 1
DiDoMi	4 | 0	2 | 2	2 | 2	0 | 4	1 | 3
Oath	5 | 0	0 | 5	0 | 5	2 | 3	0 | 5
Sourcepoint	5 | 0	5 | 0	0 | 5	0 | 5	2 | 3
Evidon	9 | 0	3 | 6	3 | 6	5 | 4	2 | 7
TrustArc	17 | 0	14 | 3	2 | 15	13 | 4	10 | 7
Quantcast	37 | 0	18 | 19	35 | 2	25 | 12	12 | 25
OneTrust	39 | 0	37 | 2	29 | 10	6 | 33	9 | 30
Custom	97 | 4	10 | 91	12 | 89	44 | 57	18 | 83
Managed	98.21%	41.25%	37.67%	43.95%	25.11%

There are two takeways could be observed from the comparison. First, although the scope of application of GDPR CMP has greatly increased in the past two years, custom solutions to GDPR consent management platform still accounts for a large proportion, which causes problems for other extensions. This is partly a problem for the extension as well, since the extension is only hardened against custom consent management in the Top500 websites. However, the solution is also able to be extended by peer sourcing, with users contributing values that work for them; this feature is not supported by other extensions to the best of the knowledge.

Second, in contrast with some extensions such as Minimal Consent, the extension does not rely solely on inspecting the HTML DOM to identify GDPR CMPs. Examining HTML DOMs to identify, for example, the modal dialog boxes asking for consents, is

Table 8.7: Comparison among the 91 **global** websites that have GDPR consents. (s|f) refers to the number of websites that are supported/failed to support. **(CC: CookieCutter, MC: Minimal Consent, COM: Consent-O-Matic, CM: Consent Manager, IDC: I Don't Care About Cookies)**

CMP	CC (s|f)	MC (s|f)	COM (s|f)	CM (s|f)	IDC (s|f)
Admiral	1 | 0	0 | 1	0 | 1	1 | 0	0 | 1
Vivendi	1 | 0	0 | 1	0 | 1	0 | 1	0 | 1
Oath	2 | 0	0 | 2	0 | 2	0 | 2	0 | 2
DiDoMi	6 | 0	1 | 5	1 | 5	2 | 4	1 | 5
TrustArc	6 | 0	0 | 6	0 | 6	2 | 4	1 | 5
Evidon	8 | 0	2 | 6	1 | 7	3 | 5	1 | 7
Quantcast	9 | 0	5 | 4	0 | 9	6 | 3	4 | 5
OneTrust	20 | 0	18 | 2	7 | 13	2 | 18	9 | 11
Custom	37 | 1	5 | 33	2 | 36	15 | 23	5 | 33
Managed	98.90%	34.07%	14.29%	34.07%	23.08%



Quantcast

We value your privacy

We and our partners store and/or access information on a device, such as cookies and process personal data, such as unique identifiers and standard information sent by a device for personalised ads and content, ad and content measurement, and audience insights, as well as to develop and improve products.

With your permission we and our partners may use precise geolocation data and identification through device scanning. You may click to consent to our and our partners' processing as described above. Alternatively you may click to refuse to consent or access more detailed information and change your preferences before consenting. Please note that some processing of your personal data may not require your consent, but you have a right to object to such processing. Your preferences will apply to this website only. You can change your preferences at any time by returning to this site or visit our privacy policy.

We value your privacy

We and our partners store and/or access information on a device, such as cookies and process personal data, such as unique identifiers and standard information sent by a device for personalised ads and content, ad and content measurement, and audience insights, as well as to develop and improve products.

With your permission we and our partners may use precise geolocation data and identification through device scanning. You may click to consent to our and our partners' processing as described above. Alternatively you may access more detailed information and change your preferences before consenting or to refuse consenting. Please note that some processing of your personal data may not require your consent, but you have a right to object to such processing. Your preferences will apply to this website only. You can change your preferences at any time by returning to this site or visit our privacy policy.

(a) CMP with one-click "Disagree"

(b) CMP with only one-click "Agree"

Figure 8.3: Different UIs of Quantcast GDPR consent management platform

an attractive approach but this approach is brittle against UI changes in the CMP **e.g., the UI components might intermittently fail for unclear reasons**, or support for custom skins and branding, which are offered as additional features by several CMPs. For instance, due to the multiple choices of skins and dashboards, Quantcast is one of the most complicated GDPR CMPs to protect against. Figure 8.3 displays two of Quantcast UIs for showing GDPR custom banners, which also are the main choices for users to configure the Quantcast UI. According to the statistics of Quantcast Choice Reports [Qua21], only 9.52% of Quantcast CMP customers are configured with *Reject All* function on the CMP ("Disagree" to the consent terms on Figure 8.3(a)). Compared with that, the proportion of customers who placed Quantcast CMP displayed as Figure 8.3(b) is up to 80.95%. I expect that users of the websites deploying this second solution agree to all tracking without making any modification. Minimal Consent processes the consent to the corresponding GDPR CMP by detecting the specific property such as className in HTML elements. The difference in the number of elements with the same class name affects the particular HTML object processed by the extension. This explains why Minimal Consent failed to automatically manage many websites that used Quantcast CMP (*cf.* Table 8.6 and Table 8.7).

8.5.2 Comparison with ad blockers

Currently the most popular tool to decrease web tracking is the use of Ad Blockers. I therefore compare the performance CookieCutterwith three different popular ad blockers: uBlock Origin, Adblock Plus, and Ghostery[4], which are recorded by previous studies as the ones most widely used [FHA20]. I visit websites collected from users four times, using either CookieCutter or one of the three ad blockers in each visit. I record the number of cookies in each cookie category during each visit. The visit sets the most private GDPR consent option it knows of. With the other Ad Blockers, I assume users accept the default GDPR consents (or are outside the EU and not shown the GDPR consent banner at all), but the users benefit from the protection offered by the ad blocker.

Figure 8.4 shows the comparison. I observe the decrease in number of non-essential cookies (*i.e.,* cookies other than strictly necessary cookies) as well as in the decline of cookies that are non-beneficial to the user (strictly necessary cookies cannot be avoided; there may be some additional cookies that are categorised as "functional" and may store information that helps user experience, such as their language or UI preferences. Non-beneficial cookies are all other cookies, such as

[4] Specifically, I use Ghostery 8.5.8: https://www.ghostery.com, AdBlock Plus 3.11: https://eyeo.to/adblockplus/chrome_install/index, and uBlock Origin 1.36.2: http://tiny.cc/ublock-origin.

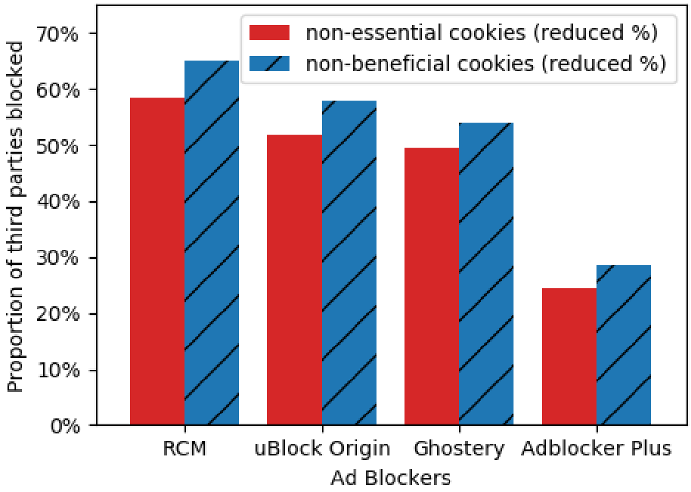


Figure 8.4: Proportion of decline in the number of (in red) non-essential third-party cookies (*i.e.*, cookies other than strictly necessary cookies) and (in blue) third-party cookies that only benefit websites and are not useful to users (*i.e.*, targeting and tracking cookies, and analytics cookies), when users apply different ad blockers to visit websites in the wild. (Larger decline is better).

targeting/advertising and analytics cookies, which only benefit the website and not the user).

CookieCutter performs better than any of the ad blockers I consider and reduces non-beneficial cookies by 65% and non-essential third-party cookiesby around 58%. The performance of uBlock Origin is second only to the extension, which also disables nearly 60% of non-essential cookies. uBlock Origin works very well for a set of websites, banning all visible ads and trackers on the websites. And the second reason why it is superior to the other two add-ons is its strong privacy protection settings by default. In addition to Easy List, uBlock also applies Easy Privacy, and other filters in Peter Lowe's tracker list[MGF19] without any extra custom settings. However, it has a disadvantage of excessively restricting some useful cookies, which is likely to result in negative impacts on some normal functions. For example, the "hot articles" section of news sites might be disabled, and some GDPR consent cookies might be removed and reappear continuously.

Due to the lack of a default mechanism to entirely opt out from "non intrusive ads", AdBlock Plus provides the worst in regards to decrease in cookie numbers, but this also breaks fewer websites compared to other Ad Blockers: AdBlock Plus is more inclined to provide users with the option to determine the visibility of each DOM in the website, i.e., hiding the element via the CSS property visibility rather than completely blocking trackers. This more moderate approach allows for some ads, which ultimately leads to weaker protection of users in terms of the number of third parties. Note that unlike Ad Blockers, CookieCutter does not meddle with the HTML DOM and works within the GDPR consent framework, so website experience is not expected to change compared to users who do not deploy CookieCutter. **For instance, some of screen spaces are left with a black vacant bar where the ad used to be, which leads to the loss of user experience. And if users are overzealous in ad blocking, it might lead to side effects such as breaking the core functionality of the site and removing genuine content.**

8.6 Deployment and Evaluation of CookieCutter In-The-Wild

In the final phase of the evaluation I deployed the final CookieCutter in the wild in a field study. Consequently, I checked the perception of the users regarding usage of the extension in the field study. In this study, I simply released the app in official extension stores for multiple browsers (Chrome, Firefox, Edge). I also included a small optional (IRB approved) survey in the extension. The results show that CookieCutter protected user privacy in the wild while ensuring usability. In this section I will detail results from the in-the-wild deployment.

8.6.1 Field study design

Newly developed extension that I released on the official extension stores of popular browsers also worked as a medium of the field study. This field study consists of two main surveys.

Pre-usage survey: Once the user installs the extension for the first time, the extension would automatically open a html survey form hosted on the client side (survey questions are attached in Appendix A.5). This survey is optional for the users and approved by the lead author's institutional review board (IRB) (as mentioned next). I also received survey responses if the users chose to complete the survey and submit them.

The pre-usage survey included questions regarding the user perception towards general internet privacy, their own personal privacy practices and finally on how they themselves deal with GDPR cookie consent notices today. Note that this survey is taken before users interacted with the extension, giving me an ecologically valid view of their privacy perceptions. This voluntary pre-usage survey primarily measured user's perceptions of privacy. Of course, if a user chose not to answer these questions, it did not impact the functionality of the extension in any way. This pre-usage took on average of 10 minutes to complete in the pilot studies. This survey as well as the next one is translated by native speakers into multiple languages (e.g., German, Chinese) so that I can get valid responses from non-English speakers too.

Post-usage survey: I conduct the second part of the field study after a random interval between 7 to 14 days from the date of first installation. Specifically, I sent participants a question pop-up which leads them to a survey form. This post-usage survey asked four questions that can be used to confirm user satisfaction with the extension and their concern for web privacy after using CookieCutter extension (survey questions are attached in Appendix A.6). Like the pre-usage survey, this survey too was optional for the users.

8.6.2 Ethical considerations

Newly developed extension functionality, field-study as well as the participatory design involved analysing/collecting data from human subjects (e.g., survey responses and cookies). Consequently, for both the studies (which concerned obtaining feedback on CookieCutter) I extensively discussed and attempted to preserve the ethics of data collection as well as data analysis in the study. I detail ethical considerations below. All of study procedures are discussed with and approved by the lead authors Institution IRB.

Ethics of automated anonymous data collection by CookieCutter: CookieCutter follows Belmont principle [Bea08] for preserving ethics and minimise data collection. Newly developed extension starts analysing browser cookies as soon as it is installed. However, to protect user privacy, any real-time analysis (e.g., cookie classification) occurs on the client side. I ask users if they consent to sending me automatic anonymised telemetry data (the default is not sending data, giving maximum control to the users). The anonymised telemetry data is extremely useful to validate the performance of CookieCutter in the wild (stating the academic mission and ways to protect their data). Specifically, for each consented user I gathered the weekly number of third-party trackers as well as specific distinct CMP templates those users encounter while browsing. However, to protect user privacy, I only collect the cookie classification results from different websites instead of the detailed cookie set information (which might contain sensitive information like name, value, expiration date etc.). Additionally, I only collect domain names for different CMPs, but not the url parameters as they can contain sensitive information too.

Ethics of data collection in pre-usage and post-usage survey: In the pre and post-usage survey for CookieCutter I did not ask users to share sensitive information, e.g., any personally identifiable data (name, gender, age, etc.). I also did not any automated collect information that could be used to identify users like IP addresses. Furthermore, I made sure that users 1) understand the purpose of this study; 2) understand and access their rights; note that the extension still supports to have all functionality even if the user decides not to take part in the survey or data collection plan.

8.6.3 Participants

Over a deployment of 12 weeks (April 12, 2021 to July 05, 2021), CookieCutter is installed by a total of 61 users (from extension store statistics). Out of them total of

45 users decided to take the pre-usage survey and 22 users responded to post-usage survey. They also opted for sharing their anonymous browsing data with me. Out of these users 43 users are Chrome users and 2 are Firefox users.

Location of users: Users in the wild has a wide country-wise spread (from total of 18 countries) with *users from multiple countries outside of GDPR jurisdiction.* I have users from GDPR-protected countries like United Kingdom, Italy, Spain, Germany, Denmark, Spain, Germany, France, Belgium, Austria, Sweden, and Hungary. However, I also have users from countries like United States, China, Brazil, India and Guatemala from outside of GDPR jurisdiction.

Table 8.8: The privacy perceptions and browsing behaviours of users from the field study.

Questions	**Responses (#users)**				
How long have you been using the Internet regularly?	1 year (5)	3-4 years (11)	5-10 years (20)	10 years (45)	
Do you major/work in CS/IT related fields?	yes (35)	no (41)	not to answer (5)		
Amount of computer time averagely spend per day	1 hour (0)	1-5 hours (16)	5-10 hours (37)	10-15 hours (26)	15 hours (2)
Do you have any ad blockers installed?	Ghostery (28)	AdBlock (32)	uBlock Origin (25)	AdGuard (11)	Privacy Badger (5)
Do you use any privacy oriented browsers?	Firefox Focus (7)	Opera (10)	Brave (19)	Tor (14)	
Do you use any privacy oriented search engine?	Duckduckgo (39)	Qwant (5)	Startpage (7)	Swisscows (2)	Ecosia (5)
Do you make sure to check privacy policies of sites visited?	always (33)	rarely (29)	crucial sites (19)		
Actions when seeing a GDPR cookie consent banner	defaults (31)	selection (43)	ignore (5)		
I get frustrated with difficulty of choosing private GDPR cookie consent options.	Strongly agree (53)	Agree (15)	Neutral (11)	Disagree (5)	Strongly disagree (3)
Are you tired to choosing cookie options on sites visited?	Yes (58)	No (23)			

Internet usage experience of users: Table 8.8 presents the overall browsing experience of 81 users from the field study as well as their privacy perceptions. 65 out of 81 of users are using internet regularly for more than 5 years and only 4 users are using the internet regularly for less than a year. Interestingly, only 35 out of 81 users (43.21%)

holds a degree and/or a job in CS/IT or related field, underlining the normalcy of the technical expertise of the population. However, even then the users spend a significant amount of time using computer—65 users (80.25%) of the users spend 5 hours or more on computer. In summary, the userbase are heavy computer users quite accustomed with internet, although many of them might not have traditional CS/IT training.

Privacy behaviours and expectations of the users: 58 users self-reported to have at least one ad blocker installed with AdBlock being the most common extension to block ads and 39 users leveraged privacy focused browsers or search engine. Interestingly, 33 users (40.75%) mentioned that they always check the privacy policies for the visited websites. This findings underline that users are pro-actively taking steps to protect their privacy in internet. **However, it also indicates the demographic limitations of this study that participants first self-selected by searching for and installing the developed browser extension. The generality would be the future direction of this work.**

Experiences with GDPR cookie consent notices: However, even then 38.27% of the users told when they encounter GDPR cookie consent notices, they just chose default. In effect 68 (84.95%) of the users agreed that they get frustrated while detecting most privacy preserving options in GDPR and 71.60% comment they are tired in that exercise. Thus, the pre-usage survey feedback from users indicate that automated assistance to manage GDPR consent cookies might help them.

Since, this need for automated assistance might be addressed by CookieCutter, I check if and how the user perception changed regarding the extension as contrasted by the responses from pre and post-usage survey.

Table 8.9: Comparison of perceptions regarding CookieCutter in pre- and post-usage surveys

Questions	% users				
	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Pre-usage (81)					
CookieCutter effectively control my data.	42.0 (34)	49.4(40)	7.4 (6)	1.2(1)	0.0 (0)
CookieCutter helps to choose my consents in a fine-grained manner.	24.7 (20)	40.7 (33)	23.5 (19)	6.2 (5)	4.9 (4)
I am concerned if CookieCutter might bloat my browser and make it run slower.	6.2 (5)	25.9 (21)	29.6 (24)	29.6 (24)	7.4 (6)
I am Concerned if CookieCutter might itself leak information.	7.4 (6)	29.6 (24)	38.3 (31)	19.8 (16)	4.9(4)
Post-usage (40)					
CookieCutter can effectively help control my data.	57.5(23)	35(14)	7.5(3)	0.0	0.0
CookieCutter helps to choose my consents in a fine-grained manner.	52.5(21)	27.5(11)	17.5(7)	0.0	0.0
I am concerned if CookieCutter might bloat my browser and make it run slower.	7.5 (3)	15(6)	40(16)	20(8)	15(6)
I am concerned if CookieCutter might itself leak information.	0.5 (2)	17.5(7)	25(10)	32.5(13)	17.5(7)

8.6.4 Improvement of user perceptions regarding CookieCutter in pre- and post-usage survey

I enquired about users' perceptions regarding CookieCutter before and after using it (in both of the pre and post usage surveys). Note that in the pre-usage survey, users opinion regarding CookieCutter would have been formed only by reading the description in the extension stores.

Table 8.9 presents the result. In pre-usage phase, 42% users believed cookies will help them protect privacy and only 24.7% strongly agreed that it will effectively choose cookie consent. In fact, 37.8% believed that CookieCutter might make their browsers slower and 35.5% believed CookieCutter might leak their sensitive cookie information.

In contrast, in post-usage phase the user perception improved significantly. After using CookieCutter for only 7 to 14 days, users 57.5% users strongly agreed CookieCutter helped them control their data privacy and 52.5% strongly agreed that the extension effectively set their cookie consent. Only 7.5% strongly agreed that CookieCutter slowed their computer and no users strongly agreed that CookieCutter leaked sensitive data while they used the extension.

8.6.5 Utility of CookieCutter to deal with cookie consent notices in the wild

Recall that CookieCutter also collected anonymous CMP template data that the consented users encountered. I analyzed these CMP templates to investigate a simple question—*What percentage of CMP templates encountered in the wild can CookieCutter effectively assist users with?*.

Section 8.4.2 already described different heuristics that I deployed to set the correct cookie values—hand crafting cookie values for popular CMP/websites, and peer sourcing. For finding answer to the question above, I further divide these CMPs according to the strategy CookieCutter took to identify GDPR cookies and set the cookie values.

Characterising CMPs encountered by CookieCutter in the field study: In the field study, CookieCutter encountered a total of 1,869 first-party websites. Among them CookieCutter detected 1,165 (62.3%) accesses to first-party websites with known CMPs. I divide the CMPS into three broad categories — (1) custom CMPs for specific websites (2) pre-identified popular CMPs (3) unidentified CMPs.

Handling CMPs encountered by CookieCutter in the field study by pre-stored cookie values: The result is shown in Figure 8.5. CookieCutter encountered 11.1% (93 websites) cases where CookieCutter used pre-stored consents. These corresponds

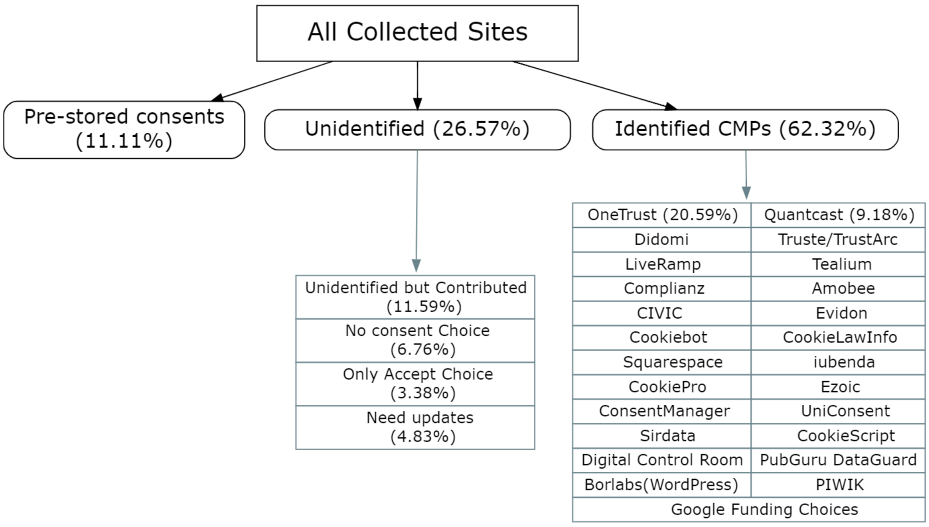


Figure 8.5: Strategies adopted by CookieCutter in the wild to assist users by finding CMPs and setting correct cookie values to protect privacy.

to custom–developed (home grown) CMPs by websites in Alexa top500, for which I have handcrafted the right values in CookieCutter. Furthermore, as I mentioned earlier, I also handcrafted and added privacy-preserving cookie values for 25 most popular CMPs (obtained from Alexa top 500 websites) in CookieCutter. They account for 62.3% of all the websites seen across the userbase. In factm two popular CMPs OneTrust and Quantcast accounts for 20.6% and 9.2% websites respectively. In summary, handcrafted values of GDPR consent cookies helped for 73.4% of all websites.

Handling unidentified CMPs encountered by CookieCutter in the field study: Out of the 26.6% unknown CMPs encountered by CookieCutter I realized that 6.8% of the cases did not have a GDPR consent banner and 3.4% gave “only accept” choice (i.e., mentioned “if you use the website you are agreeing to being tracked”). Clearly, an automated method designed to assist used with GDPR protection cannot help in these cases—these websites are actively trying to circumvent GDPR protection. However, CookieCutter helped for rest of the websites with unidentified CMPs.

For a significant 11.6% of sites values of GDPR consent cookies for unknown CMPs (not pre-stored) were contributed by peer-sourcing. Note that, these values were set manually by some users and then shared in CookieCutter so that other users can reap benefit of their effort. Superficially, for these 11.6% cases, the extension can detect the name of the GDPR consent cookies but does not know what value to set for those cookies—peer-sourcing helped to set the value of these cookies. Finally, for only 4.8% of the websites CookieCutter encountered an unknown CMP but could not help users with choosing the correct cookie values.

Summary: Pre-stored hand-crafted cookie values included in CookieCutter helped set the values of GDPR consent cookies for 73.4% of websites in the wild. Furthermore, for a non-trivial 11.6% sites values of GDPR consent cookies were set using peer sourcing even in the small-scale field study. Overall, the field study shows that the combination of pre-stored CMP values and peer-sourcing is already enabling CookieCutter to automatically assist users with 85% of all the websites that they visit.

8.6.6 Efficacy of user contributed values of GDPR consent cookies in CookieCutter

More than 11% of visited websites CookieCutter set the values of GDPR consent cookies using peer sourcing. Thus, I next investigated how well the user contributed values of GDPR consent cookies work in the field study to assist CookieCutter users to choose privacy preserving cookie consent.

Measuring effectiveness of the values of user contributed GDPR consent cookies: Out of the 11.6% websites where CookieCutter used peer-sourcing, in total, the extension encountered unique 65 top-level domains. For each of these domains more than 3 users contributed values of their GDPR consent cookies. In order to measure the effectiveness of these GDPR consent cookies contributed values, I did the following simple experiment. I loaded these particular cookie values in an automated browser and measured the number of different type of cookies stored in the browser before and after storing values of these GDPR consent cookies for a given domain. Naturally, the measurement of effectiveness for values of these GDPR consent cookies is guided by whether they are in-effect reducing the number of different types of cookies a website is storing compared to the default "accept all" setting.

Table 8.10: Total reduction of number of cookies while setting user-contributed GDPR consent cookies for particular domains. **In order to detect the ability to minimise risks and potential vulnerabilities,** the baseline of comparison are the number of cookies set when a user chose default accept settings **i.e. the risk bounds of cookies**.

Cookie type	Most Decline	Least Decline
Strictly Necessary	27.64%	20.41%
Functional	43.18%	34.94%
Performance	41.75%	32.89%
Targeting/Advertising	**54.04%**	**44.61%**

Reduction of number of cookies by user-contributed GDPR consent cookies: For 36 domains (out of 65 domains), user contributed GDPR consent cookies effectively reduced the number of Performance and Targeting/Advertising cookies to zero. Table 8.10 presents more detailed results on effectiveness of user contributed GDPR consent cookies. Note that, for same domain I collected multiple user-contributed GDPR consent cookies, thus creating the most and least decline percentages. The table demonstrates that, user contributed GDPR consent cookies enables the avoidance of non-essential third-party cookies via CookieCutter. Furthermore, even the least effective user contributed GDPR consent cookies reduced the number of Targeting/Advertising by 44.6%, underlining the efficacy of peer sourcing mechanisms of CookieCutter in a real-world deployment.

Summary: In summary, user provided values of GDPR consent cookies helped other CookieCutter users in the field study for 11.6% of the websites. The exploration demonstrated that these user-contributed GDPR consent cookies indeed helped to users to choose restrictive cookie consent settings which significantly decreased the number of non-essential cookies.

8.7 Conclusion

In this work, I introduce CookieCutter, a system to protect users by automatically detecting GDPR consent cookies and setting their values. The participatory design as well as field study demonstrated the utility of CookieCutter in protecting user privacy. However, I believe CookieCutter is only the first step that paves a way for future research on using automated techniques for GDPR enforcement. For example, two current limitations for CookieCutter are — First, currently CMPs can easily decide to allow tracking globally by adding location check and tracking consent check before setting tracking cookies to circumvent CookieCutter. Hardening against such attacks is a concrete future work. Second, CookieCutter still does not solve the problem of those sites which do not have cookie consent notices or simply inform the user that by using the website they agree to being tracked [HS19]. Identifying these sites and creating a synergy of legal as well as technical solutions is another exciting future direction in this research domain of assisting GDPR enforcement to help users. However, I believe CookieCutter demonstrates the potential of such techniques and the practical utility of deployment for automated tools to assist users in protecting their privacy while browsing.

Chapter 9

Conclusions and Future work

Summary

In general, there are composable design patterns for the browser privacy protection: First, when seeking applications for properties of the browsing/online activity network, I make efforts to understand each party's properties to characterise the third-party ecosystem and identify the natures. It is to describe the relationshiip between the interest of users and tracking status both experimental and real environments. Then, the relevant cookie classification based on the preliminary analysis would be incorporated into the automated framework, which alleviates the requirements for user-side expertise in a given application domain. Combined with the principle of the“data minimization”, this might contribute to the future implementation of automatic maximisation of user privacy in browsers.

9.1 Conclusion

The principle for this dissertation has been understanding how users' privacy among browsing histories can be leveraged by third parties for commercial purposes. Given the characteristics of the third-party ecosystem, it has become a double-edged sword. It allows advertisers to achieve precise positioning of users and push personalised advertisements with the help of cookies; however, for users, a huge amount of private information is at risk of leakage.

Considering that the deprecation of third-party cookie ecosystem might severely disrupt the structure of online communities and stifle the survival space of startups and emerging companies in advertising, exploring the approach to protect users with a standardised application model and ensure that users could experience the website necessary services while being protected has become the focus of this thesis.

This dissertation explores two realms of this problem, one working with Alexa top websites, and other looking at a vast set of real users acound the world. In both cases, the key was the set of methods used to tease out the characterics of third-party ecosystem. According to the status quo of digitisation, business development might be highly contingent upon online data, which makes the browsing data stored in third-party ecosystems valuable. **The work done in my dissertation depicts thatthe effect of data protection laws that regulators are committed to promoting for data-driven businesses, in addition to privacy protection tools such as browser extensions.** However, inspired by the release of data privacy regulations (such as GDPR, CCPA, LGPD, POPI, etc.), users' dark pattern of winning "informed consent" came into being. Legally, vendors are required to obtain "informed consent" from users before any processing of non-essential cookies. The dark pattern would affect the decision of users through the special settings of the options placed in the cookie banners

In response to issues stated above, I first characterise the third-party cookie ecosystem across the countries in §4. **It illustrates that the fixed user group across time and the specialisation interests is prone to be tracked. Furthermore, the nature of country-specific trackers determines the structure of the country's cookie ecosystem, which results in the intensity of domestic user data protection demands. Therefore, trackers would rely on the traffic discrimination to some extent while loading third parties over time, segmenting it according to the connection location and the browsing preference (i.e. the category of the website that the user prefers to browse).**

Then, relying on the automatic browser container allocation of contextual identities (§5) and the expansion of the existing cookie classification database**, it is possible to validly protect the privacy**. The former **characterises and quantifies the connection status of third parties that serve a given set of first-party websites, named "Tangle Factor",** realising the isolation of inner-connected websites through the hidden third parties. **In a sense, it fills in the gap of analysis of hidden inner-connectivity among first-party websites in the traditional model, and the research demonstrates a positive correlation between the value of the Tangle Factor and the interconnectedness of third party cookie ecosystem.**

And the latter focuses on the accurate interception of stored attributes by cookies (§6). **Since the majority of users are unwilling to carefully consider the complicated implications of sharing data, it is likely to lead to the "Reject All" decision. The issue is that some strict settings for cookies classification in specific sites might affect the user's experience of normal site functions under the principle of data minimisation (i.e., "Reject All"). The capability requirements for GDPR-like regulations have further accelerated user demand for browser extensions.** Therefore, to simplify the decision-making stage and ensure customised user privacy, Chapter 8] explore the automatic consent management based on the GDPR compliance framework**, ensuring that continuous pop-ups of cookie banners would not interrupt the user experience. Furthermore, I set up a peer sourcing mechanism to guarantee the feasibility of more shared consents, balancing "consent fatigue" and "user privacy".**

9.2 Open problems

The journey of the dissertation is a complete long procedure comprising on various stages. There are some interesting problems come across in the journey of the dissertation but missed due to the restricted time, arena and interest of focus. At the least, I would like to enumerate some of them for the further investigation and discussion.

In the era of the cookie ecosystem of the browser, it is still vital to build fully automated assistance systems with a universality for different levels of users. **It is an interesting open problem as to whether, the knowledge and willingness of participants to download the browser extension leads to some biased cohort towards a specific demographic segmentation to some extent. In particular, all preliminary users sampled have computer-related jobs and are young adults. On another note, cookies are not the only way to track users. Trackers might use other technologies such as fingerprinting the browser of the user for canvas recognition. It might result in an under-approximation of the magnitude of the problem. However, it possibly makes the study prone to false positives [AFM20, LZC16]. Therefore, a more elaborate way of profiling users when computing the third-party overlaps is still open.**

As it stands, the privacy protection tools are developed independently against specific issues, to address the individual problems at hand (i.e., browsing history, GDPR dark patterns, etc.). Based on that, we pave a path forward for the exploration of demographics for more general user base to happen. Moving forward, I foresee a unified and compliant browser privacy protection framework that assists normal users in professional privacy protection without affecting the original service and information

The compliant browser privacy protection framework is responsible for identifying user-friendly services and profit-driven trackers, preventing non-professional users from the redundant and misleading messages displayed by vendors. **Noted that the key to the framework is to ensure the availability of complete third-party functionality and user privacy. My work done on the "container” allocation design, allowed me to appropriately isolate third parties. One of the natural extensoins would be supported by the interests identification of users and push the relevant advertisements or other targeted services based on the interest-based recommendations, thus the "container” service could guarantee the categorisation-based customisation under GDPR compliance.**

Thirdly, the user could customise the acceptance range of third-party vendors. For instance, when visiting a shopping website, only vendors providing shopping, sports, reaction-related services are accessible. **The possibility of linking privacy metrics in first-party websites, with affiliated third-party trackers is what I define as customised tracking driven service. This aspect of third-party service mode could actually be immensely helpful in developing user-side privacy interventions with minimum costs but maximum balance.** The framework would allocate the site to the proper "container" based on the user's customisation. The dynamic allocation also avoids the potential service interruption or risk of being affected.

9.3 Future Work

Finally, I would like to introduce a few aspects of the path forward for future work.

First, based on the analysis of third-party ecosystems done, it could be extended to a wider range of countries, studying the effect that other demographic features (e.g., the language, age or gender of users) might have in the strategies used to track users. Spending the time in the exploration of the trade-off between usability and privacy, for example, having an OAuth account in the right container. Except that, the proposed categorisation could be taken advantage in a multi-country context to systematically explore how to cluster websites into containers in a private manner.

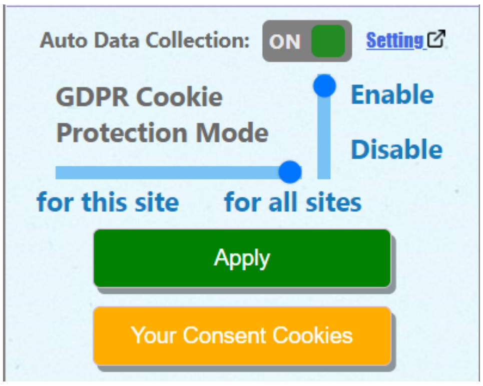
Secondly, current consent management systems based on the GDPR compliance framework is weak in managing the cookie classification, and some manual operations lead to the efficiency loss due to the uncompensable absence of cookie consent notifications. Therefore, a novel usable browser extension is on the promotion, which assists users in setting the current values of GDPR consent cookies that '**targets**' data minimisation. **the scope of the novel framework is more flexible than the analysis scope of the tracker, which adjustably unify the cross-site privacy decisions. Alternatively, the classification could guide the appropriate exposure level based on the third-party interconnectivity.** Ideally, it could identify sites that do not have cookie consent notices or simply inform the user that by using the website they agree to be tracked. **And this interesting topics could create a synergy of legal as well as technical solutions, exploiting GDPR enforcement of cookie consents to assist in user privacy protection.**

Appendix A

Supplementary materials of CookieCutter

A.1 Preliminary version of interface

The old version of CookieCutter displays in Figure A.1. In the old version, we initially gave users the right to choose whether to automatically apply protection to all websites.The vertical slider is used to disable individual website and the parallel one is for the coverage of protection.



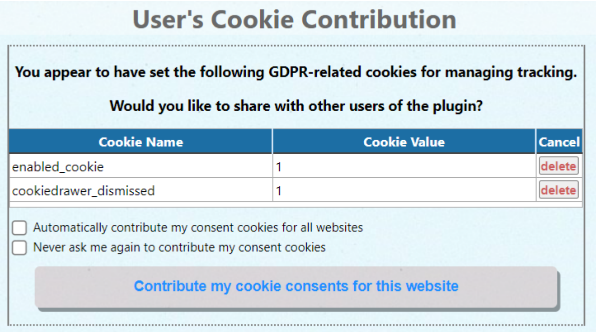
Auto Data Collection: ON

Figure A.1: Pre-version of extension auto protection UI.

A.2 User-contributed interface

In order to interactively collect GDPR consent cookies from users, we would create an interface when users are willing to contribute customised GDPR consent cookies.

Figure A.2 is the newly created tab on the window for users, displaying the details of GDPR consent cookies (names and values). If the user is unwilling to send any of GDPR consent cookies, the delete button allows the user to abandon sending the cookie in a particular line.



User's Cookie Contribution

You appear to have set the following GDPR-related cookies for managing tracking.

Would you like to share with other users of the plugin?

Cookie Name	Cookie Value	Cancel
enabled_cookie	1	delete
cookiedrawer_dismissed	1	delete

□ Automatically contribute my consent cookies for all websites
□ Never ask me again to contribute my consent cookies

Figure A.2: User-contributed interface in the newly created tab.

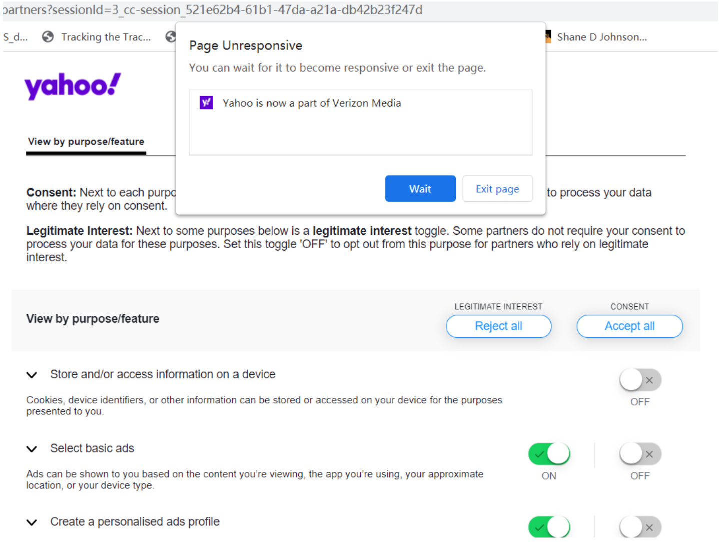
A.3 Problem caused by unresponsive CMP

Some of websites use the separate page for GDPR consent management, which raises another problem. Since the user's first visit would redirect the user to the consent page (Figure A.3), the unexpected crash of the page would prevent users from accessing the website. Unlike the notification banner placed at the bottom or top of the website, this kind of consent page could not be ignored and completely affect the user experience.

A.4 Participatory design interview script

In our extension, we have two main panels and one toolbar for you to examine and give us feedback about. We will ask you some questions regarding each of them after presenting them. Feel free to ask any clarification regarding the extension components during the interview.

In the "Main Panels" We will provide cookie information for each of your currently opened tab as well as controlling consent for those cookies=. We will also give statisti-



partners?sessionId=3_cc-session_521e62b4-61b1-47da-a21a-db42b23f247d

S_d... Tracking the Trac... Shane D Johnson...

Page Unresponsive

You can wait for it to become responsive or exit the page.

Yahoo is now a part of Verizon Media

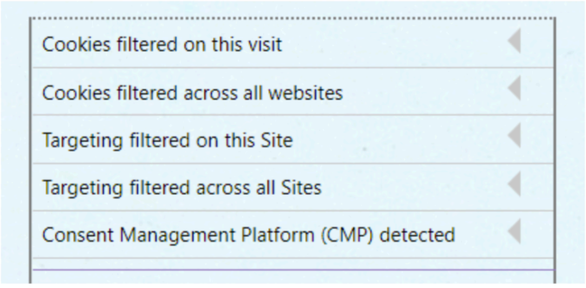
Consent: Next to each purpo to process your data where they rely on consent.

Legitimate Interest: Next to some purposes below is a **legitimate interest** toggle. Some partners do not require your consent to process your data for these purposes. Set this toggle 'OFF' to opt out from this purpose for partners who rely on legitimate interest.

View by purpose/feature	LEGITIMATE INTEREST	CONSENT
	Reject all	Accept all
Store and/or access information on a device Cookies, device identifiers, or other information can be stored or accessed on your device for the purposes presented to you.		OFF
Select basic ads Ads can be shown to you based on the content you're viewing, the app you're using, your approximate location, or your device type.	ON	OFF
Create a personalised ads profile	✓	✕

Figure A.3: An example of crashed CMP consent page of Yahoo.

cal information about your browsing history since you installed our extension in the toolbars.



Cookies filtered on this visit

Cookies filtered across all websites

Targeting filtered on this Site

Targeting filtered across all Sites

Consent Management Platform (CMP) detected

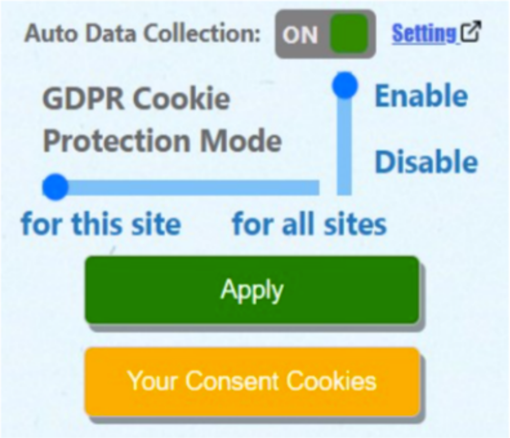
Figure A.4: First main panel of browser extension popup shown to participants.

Answer the questions after checking the first main panel (Figure A.4).

1. Can you smoothly open each collapsible table row for detailed information?
2. What do you think the table is showing you? (Compared with our explanation above)
3. Does it show the number of cookies/ local storage / session storage correctly?
4. When it detects a specific CMP for the currently visited website, can you successfully manage the cookie setting with one click of "Apply"? And the cookie banner will not bother you anymore?
5. Are there any websites that require you to provide custom cookie settings that we fail to provide support? Please text it for further improvement.

6. Do you think these 5 rows have provided all the essential information that you are interested in?

7. Is there anything you would want to know but isn't explained on this part?



Auto Data Collection: ON

Figure A.5: Second main panel of browser extension popup shown to participants.

Answer the questions after checking the second main panel of extension popup (Figure A.5).

1. Do you fully understand how to stop your consent data from collection (by toggle switch)?

2. What do you think "Enable" and "Disable" the auto consent management does?

3. Do you prefer automatically running CMP detection and consent settings in the background each time you open a new tab? Or, is it preferable that it should only work when you open the extension tab for help?

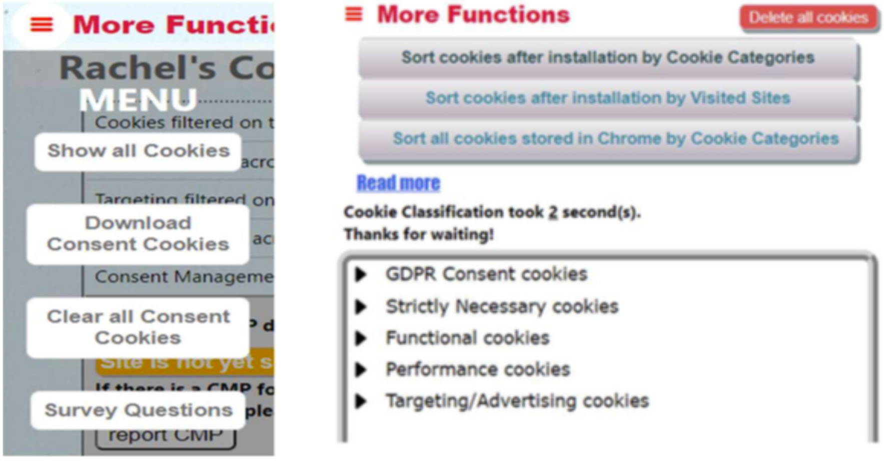
4. What do you think would happen if you set auto cookie protection mode "for all sites" and "Apply"?

5. What do you think would happen if you set a website as "Disabled"?

6. If you have set a website to "Disabled" and visit again, will it still remain in your setting as "Disabled"?

7. Is there anything you would want to know but is not explained by hyperlinks on this tab?

8. Is there any button or text which is confusing to you? (Change/add/remove to make it clearer?)



Cookie Classification took 2 second(s).
Thanks for waiting!

(a) Toolbar 1

(b) Toolbar 2

Figure A.6: Main toolbars of our extension popup.

Now we are showing your parts of a toolbar included in our extension (in Figure A.6
).

1. Do you think the Download function makes sense to you?
2. Do all of the buttons work well?
3. What is confusing on this toolbar to you? Any buttons?
4. Are there any functions you would like to add?
5. What do you think the section"Sort cookies after installation by cookie categories" is showing you?
6. After you have accessed several websites, If you click on "Sort cookies after installation by visited sites", can you view the cookie categorized by first-party websites(visited domains)?
7. Do you think there is any difference between the first and third options?
8. Do you think the classification makes sense to you?
9. When you have visited 1k+ sites, do you feel that this tab has become slower? Is the extent of slowing acceptable?
10. Do you feel like this tab is telling you new and interesting information, compared to the main panels?
11. Would you add/change/remove anything to make this panel more interesting?

12. Would you go back to the previous panel after you viewed the “More Function” panel?
13. Is there anything you would want to know but is not included on this panel?

Finally Please answer the following questions:

1. On a scale of 1 to 5, with 1 being the least and 5 being the most, how informative would you say the extension was?
2. On a scale of 1 to 5, how knowledgeable would you say you are now, after this session, about how GDPR consent management and cookie works?
3. On a scale of 1 to 5, how interested would you say that you are now, after this session, in learning more GDPR consent management and cookie classification?

A.5 Pre-usage survey for our field study with privacy-oriented questions

We assign the scores for each option in our privacy-oriented questions on a scale of 1 to 5. We use black fonts to display the detailed questions and corresponding options, red fonts in the brackets represent scores.

Understanding your stance on privacy

1. I believe companies seeking information online should have a detailed online privacy policy (clearly disclosing the way the data are collected, processed, and used).

   ○ Strongly agree (5”) ○ Agree (4”) ○ Neutral (3”) ○ Disagree (2”) ○ Strongly disagree (1”)

2. I should be made aware of how my personal information will be used.

   ○ Strongly agree (5”) ○ Agree (4”) ○ Neutral (3”) ○ Disagree (2”) ○ Strongly disagree (1”)

3. It usually bothers me when online companies ask me for personal information.

   ○ Strongly agree (5”) ○ Agree (4”) ○ Neutral (3”) ○ Disagree (2”) ○ Strongly disagree (1”)

4. I am concerned that online companies that collected personal data is sharing it with other companies I don’t know about.

   ○ Strongly agree (5”) ○ Agree (4”) ○ Neutral (3”) ○ Disagree (2”) ○ Strongly disagree (1”)

5. I find that online ads are sometimes useful and I dont mind being tracked if it gives new and useful information.

   ○ Strongly agree (1”) ○ Agree (2”) ○ Neutral (3”) ○ Disagree (4”) ○ Strongly disagree (5”)

Privacy oriented actions

1. Do you clear your Internet browser history regularly?

   (a) No, I do bother clearing my browser history. (1”)

   (b) No, I don’t manually clear my browser history, but I set my browser to clear at regular intervals. (2.3”)

   (c) Yes, I manually clear my browser history at short intervals (e.g., every day/week or month). (3.7”)

   (d) Yes, I occasionally clear my browser history (no regular intervals). (5”)

2. Do you clear your browser cookies regularly?

   (a) No, I do bother clearing my cookies. (1”)

   (b) No, I don’t manually clear my cookies, but I set my browser to clear at regular intervals. (2.3”)

   (c) Yes, I manually clear my cookies at short intervals (e.g., every day/week or month). (3.7”)

   (d) Yes, I occasionally clear my cookies (no regular intervals). (5”)

3. Do you have any of the following ad blockers installed? (have any one: 5”; not have: 1”)

   □ Ghostery □ AdBlock (Plus) □ uBlock Origin □ AdGuard □ Privacy Badger

4. Do you use any of the following privacy oriented browsers? (have any one: 5”; not have: 1”)

   □ Firefox Focus □ Opera □ Brave □ Tor

5. Do you use any of the following privacy oriented search engine? (have any one: 5”; not have: 1”)

   □ Duckduckgo □ Qwant □ Startpage □ Swisscows □ Ecosia

Deal with GDPR cookie consents

1. Do you make sure to check the privacy policies of the websites you visit?

   (a) Yes, I always check (5”)

   (b) I make sure to check for important sites like banks or when I share credit card information (3”)

   (c) I rarely check (1”)

2. When you see a GDPR cookie consent banner, what is your usual action.

   (a) go through the banner and choose the most privacy oriented setting (5”)

   (b) accept default (5”)

   (c) carry on with the cookie banner showing (1”)

3. I get frustrated or angry when websites make it difficult to choose a privacy oriented option.

   ○ Strongly agree (5”) ○ Agree (4”) ○ Neutral (3”) ○ Disagree (2”) ○ Strongly disagree (1”)

4. Do you feel tired by having to choose all the different cookie consent options on all the sites you visit?

   (a) Yes (5”)

   (b) No (1”)

A.6 Post-usage survey questions

In the post-usage survey, we ask users about their experience and satisfaction with CookieCutteron a scale of 1 to 5. After installing CookieCutter1-2 weeks (random intervals between 7 to 14 days), this survey would pop up on the main panel to the user.

1. This extension effectively puts me in control of my data.

   ○ Strongly agree ○ Agree ○ Neutral ○ Disagree ○ Strongly disagree

2. After installing this extension, I miss being able to choose my cookie consents in a fine-grained manner.

   ○ Strongly agree ○ Agree ○ Neutral ○ Disagree ○ Strongly disagree

3. I am worried that this extension might bloat my browser and make it run slower.

   ○ Strongly agree ○ Agree ○ Neutral ○ Disagree ○ Strongly disagree

4. I am worried that this extension might itself leak information.

   ○ Strongly agree ○ Agree ○ Neutral ○ Disagree ○ Strongly disagree

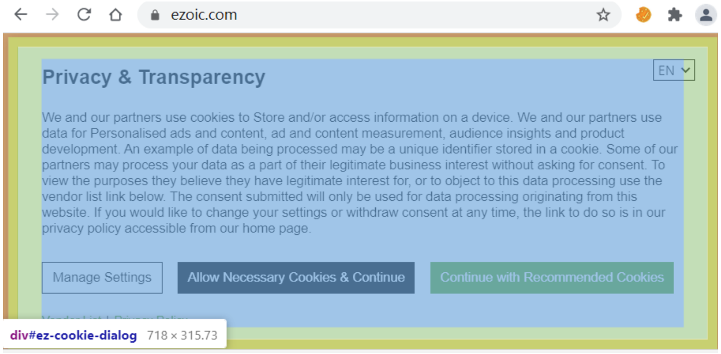
A.7 Regular expression of GDPR consent cookies

In Table A.1, there displays the regular expression matching GDPR consent cookies we used for CookieCutter. Cookie names would be converted to lowercase before *GDPR regex*.

Table A.1: Regular Expression matching

<table>
<tr><th></th><th><i>Regular Expression</i></th></tr>
<tr><td>CookieNames (Lowercase)</td><td>/uniconsent-|li_gc|admrla|^optanon|euconsent|gdpr|pref-agent|cookie-?[preferences?|notice|settings?|status]|consent|ba_cookies|_sp_v1_|et_cookies|_gali|ckns_|cookie[drawer]_dismissed|_privacy_|oil_gtm_cookie|hasseencookie(disclosure)?|policy(-update)?-notice|notice_poptime|user_prefs|preferencesmsn|^cpol$|cookies_[advertising|analytics|functional|settings|notice]|cookieinfo$|eu_cookie$|cookie_notif|[cbc|tracking]-cookie-status|cookielaw|cookie-notice-donottrack|^ccp$|performance|preferences?|marketing|kvkk_notificatioon|eucookielaw|noticeclosed|aep_usuc_|googlepersonalization|care_about_cookie|privacyprompt|oup-cookie|^pi_opt_in|^mca_vid$|privacyversion|borlabs-|ppms_privacy_|compliancecookie|eupubconsent|jg[advertising|analytics|functional|settings|notice]|^dcbn$|storagessc|bkng$|didomi_token|usprivacy|hasseennotice$|cmapi_cookie/g</td></tr>
</table>

A.8 CMP detection in HTML



Privacy & Transparency

We and our partners use cookies to Store and/or access information on a device. We and our partners use data for Personalised ads and content, ad and content measurement, audience insights and product development. An example of data being processed may be a unique identifier stored in a cookie. Some of our partners may process your data as a part of their legitimate business interest without asking for consent. To view the purposes they believe they have legitimate interest for, or to object to this data processing use the vendor list link below. The consent submitted will only be used for data processing originating from this website. If you would like to change your settings or withdraw consent at any time, the link to do so is in our privacy policy accessible from our home page.

Manage Settings
Allow Necessary Cookies & Continue
Continue with Recommended Cookies

div#ez-cookie-dialog 718 × 315.73

Figure A.7: HTML DOM Source of GDPR consent management platform template

There is a screenshot showing how to detect the DOM of the GDPR CMP embedded in a website. The specific CMP would use the fixed the className or id property (Figure A.7). Therefore, we have a match-up table of CMP and the corresponding DIV class Name or id, obtaining the corresponding CMP version after capturing the DOMs of the website currently visited by the user.

Bibliography

[3018] GDPR Recital 30. Online identifiers for profiling and identification. https://gdpr-info.eu/recitals/no-30/, May 2018. Accessed on 2018-07-25.

[Adb18] Adblock_Plus. Allowing acceptable ads in adblock plus. Available at https://adblockplus.org/acceptable-ads (last accessed on 25 June 2021), 2018.

[AdG19] AdGuard. Faq-what is the difference between adguard filtering methods? Available at https://kb.adguard.com/en/android/faq#what-is-the-difference-between-adguard-filtering-methods (last accessed on 25 June 2021), 2019.

[AFM20] Nasser Mohammed Al-Fannah and Chris Mitchell. Too little too late: can we control browser fingerprinting? *Journal of Intellectual Capital*, 2020.

[AHH16] Ibrahim Altaweel, Maximillian Hils, and Chris Jay Hoofnagle. Privacy on adult websites. In *Altaweel et al., Privacy on Adult Websites, Workshop on Technology and Consumer Protection (ConPro'17), co-located with the 38th IEEE Symposium on Security and Privacy, San Jose, CA (2017)*, 2016.

[AJP+20] 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 *Proceedings of The Web Conference 2020*, pages 1479–1490, 2020.

[Alb16] Jan Philipp Albrecht. How the gdpr will change the world. *Eur. Data Prot. L. Rev.*, 2:287, 2016.

[Ale18] Alexa. Global top websites. https://www.alexa.com/topsites, May 2018. Accessed on 2018-05-20.

[Ale20a] Alexa. Global top websites by category (web archive). https://web.archive.org/web/20200721103216/www.alexa.com/topsites/category, July 2020. Accessed on 2021-04-01.

[Ale20b] Alexa. Top sites by category has been retired. https://support.alexa.com/hc/en-us/articles/360051913314, 2020. Online; accessed 03 March 2021.

[Ama18] Amazon. Ad personalization. https://www.amazon.com/adprefs, May 2018. Accessed on 2018-05-30.

[Aro18] Ananay Arora. Preventing wireless deauthentication attacks over 802.11 networks. *arXiv preprint arXiv:1901.07301*, 2018.

[Ato05] EasyList Atom. Easylist, 2005.

[AZXW19] Mshabab Alrizah, Sencun Zhu, Xinyu Xing, and Gang Wang. Errors, misunderstandings, and attacks: Analyzing the crowdsourcing process of ad-blocking systems. In *Proceedings of the Internet Measurement Conference*, pages 230–244. ACM, 2019.

[BCHT19] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. Thematic Analysis. In *Handbook of Research Methods in Health Social Sciences*, pages 843–860. Singapore, 2019.

[BCJ+14a] Lujo Bauer, Shaoying Cai, Limin Jia, Timothy Passaro, and Yuan Tian. Analyzing the dangers posed by chrome extensions. In *2014 IEEE Conference on Communications and Network Security*, pages 184–192. IEEE, 2014.

[BCJ+14b] Lujo Bauer, Shaoying Cai, Limin Jia, Timothy Passaro, and Yuan Tian. Analyzing the dangers posed by chrome extensions. In *Communications and Network Security (CNS), 2014 IEEE Conference on*, pages 184–192. IEEE, 2014.

[BCK+14] Paul Barford, Igor Canadi, Darja Krushevskaja, Qiang Ma, and Shan Muthukrishnan. Adscape: Harvesting and analyzing online display ads. In *Proceedings of the 23rd international conference on World wide web*, pages 597–608, 2014.

[BDH18] David Basin, Søren Debois, and Thomas Hildebrandt. On purpose and by necessity: compliance under the gdpr. In *International Conference on Financial Cryptography and Data Security*, pages 20–37. Springer, 2018.

[Bea08] Tom L Beauchamp. The belmont report. *The Oxford textbook of clinical research ethics*, pages 149–155, 2008.

[BEK+16] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proc. Priv. Enhancing Technol.*, 2016(4):237–254, 2016.

[Bet] Better. Trackers collections. Available at https://better.fyi/trackers/.

[Bin21] Chetna Bindra. Building a privacy-first future for web advertising. Available at https://blog.google/products/ads-commerce/2021-01-privacy-sandbox/ (last accessed on 25 June 2021), 2021.

[Bir17] Clare Birchall. *Shareveillance: The dangers of openly sharing and covertly collecting data.* U of Minnesota Press, 2017.

[BJKL04] Steven Bellman, Eric J Johnson, Stephen J Kobrin, and Gerald L Lohse. International differences in information privacy concerns: A global survey of consumers. *The Information Society*, 20(5):313–324, 2004.

[BKCB22] Dino Bollinger, Karel Kubicek, Carlos Cotrini, and David Basin. Automating cookie consent and GDPR violation detection. In *31st USENIX Security Symposium (USENIX Security 22)*, page TBA, Boston, MA, August 2022. USENIX Association.

[BLSD$^+$20] Nataliia Bielova, Arnaud Legout, Natasa Sarafijanovic-Djukic, et al. Missed by filter lists: Detecting unknown third-party trackers with invisible pixels. *Proceedings on Privacy Enhancing Technologies*, 2020(2):499–518, 2020.

[BO20] Céline Brassart Olsen. To track or not to track? employees' data privacy in the age of corporate wellness, mobile health, and gdpr. *International Data Privacy Law*, 2020.

[Bor13] Frederik Zuiderveen Borgesius. Behavioral targeting: A european legal perspective. *IEEE security & privacy*, 11(1):82–85, 2013.

[BP17] Willem Boumans and Ir Erik Poll. Web tracking and current countermeasures. 2017.

[BRAY17] Justin Brookman, Phoebe Rouge, Aaron Alva, and Christina Yeung. Cross-device tracking: Measurement and disclosures. *Proceedings on Privacy Enhancing Technologies*, 2017(2):133–148, 2017.

[Bro19] Martin Brodin. A framework for gdpr compliance for small- and medium-sized enterprises. *European Journal for Security Research*, pages 1–22, 2019.

[BW18] Muhammad Ahmad Bashir and Christo Wilson. Diffusion of user tracking data in the online advertising ecosystem. *Proceedings on Privacy Enhancing Technologies*, 2018(4):85–103, 2018.

[CABM16a] A. Cahn, S. Alfeld, P. Barford, and S. Muthukrishnan. What's in the community cookie jar? In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 567–570, 2016.

[CABM16b] 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.

[ccp] California consumer privacy act (ccpa). Available at https://oag.ca.gov/privacy/ccpa.

[CDL+18] Giuseppe Contissa, Koen Docter, Francesca Lagioia, Marco Lippi, Hans-W Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence. *Available at SSRN 3208596*, 2018.

[CDS+18] Bart Custers, Francien Dechesne, Alan M Sears, Tommaso Tani, and Simone van der Hof. A comparison of data protection legislation and policies across the eu. *Computer Law & Security Review*, 34(2):234–243, 2018.

[Cen20] China Internet Network Information Center. 2020 national research report on internet use of minors. Available at http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/qsnbg/202107/P020210720571098696248.pdf, 2020.

[CGI+20] Federico Cozza, Alfonso Guarino, Francesco Isernia, Delfina Malandrino, Antonio Rapuano, Raffaele Schiavone, and Rocco Zaccagnino.

Hybrid and lightweight detection of third party tracking: Design, implementation, and evaluation. *Computer Networks*, 167:106993, 2020.

[chi18a] China social media statistics & facts in 2018 to shape your marketing strategy. Available at https://www.marketingexpertus.co.uk/blog/china-social-media-statistics-2018 (last accessed on 25 June 2021), 2018.

[Chi18b] Victoria Chico. The impact of the general data protection regulation on health research. *British medical bulletin*, 128(1):109–118, 2018.

[Chr20] Chromium. Building a more private web: A path towards making third party cookies obsolete. Available at https://blog.chromium.org/2020/01/building-more-private-web-path-towards.html, 2020.

[CKB12] Abdelberi Chaabane, Mohamed Ali Kaafar, and Roksana Boreli. Big friend is watching you: Analyzing online social networks tracking capabilities. In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, pages 7–12, 2012.

[Cli17] Cliqz. Ghostery. https://github.com/ghostery, 2017.

[CNS20] John Cook, Rishab Nithyanand, and Zubair Shafiq. Inferring tracker-advertiser relationships in the online advertising ecosystem using header bidding. *Proceedings on Privacy Enhancing Technologies*, 2020(1):65–82, 2020.

[Col18] Cookie Collective. Five models for cookie law consent, 2018.

[Cou16] Council of European Union. Regulation (eu) 2016/679 (general data protection regulation), 2016.

[CPJ18] Hanbyul Choi, Jonghwa Park, and Yoonhyuk Jung. The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior*, 81:42–51, 2018.

[Cra13] Florian Cramer. Post-digital aesthetics. *Jeu de paume-le magazine*, 2013.

[Dai] Tim Daish. The webpage toaster third-party database. Available at https://www.webpagetoaster.com/list3pdb.php.

[DCB19] Paul Daugherty and Marc Carrel-Billiard. The post-digital era is upon us-are you ready for what's next, 2019.

[Dis] Disconnect. Disconnect third-party database. Available at https://github.com/disconnectme/disconnect-tracking-protection/blob/master/services.json.

[Dis13] Disconnect. Take back your privacy. Available at https://disconnect.me/, 2013.

[DK19] Olha Drozd and Sabrina Kirrane. I agree: Customize your personal data processing with the core user interface. In *International Conference on Trust and Privacy in Digital Business*, pages 17–32. Springer, 2019.

[DLT18] Amit Datta, Jianan Lu, and Michael Carl Tschantz. The effectiveness of privacy enhancing technologies against fingerprinting. *arXiv preprint arXiv:1812.03920*, 2018.

[DM14] Jake Drew and Tyler Moore. Automatic identification of replicated criminal websites using combined clustering. In *2014 IEEE Security and Privacy Workshops*, pages 116–123. IEEE, 2014.

[DMU+19] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. Measuring cookies and web privacy in a post-gdpr world. In *International Conference on Passive and Active Network Measurement*, pages 258–270. Springer, 2019.

[DUL+18] 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.

[DYLH20] Huixi Dong, Fangru Yang, Xiaozi Lu, and Wei Hao. Internet addiction and related psychological factors among children and adolescents in china during the coronavirus disease 2019 (covid-19) epidemic. *Frontiers in Psychiatry*, 11:751, 2020.

[Eas] EasyList. Easylist overview. Available at https://easylist.to/pages/other-supplementary-filter-lists-and-easylist-variants.html.

[EJPARHF17] José Estrada-Jiménez, Javier Parra-Arnau, Ana Rodríguez-Hoyos, and Jordi Forné. Online advertising: Analysis of privacy threats and protection approaches. *Computer Communications*, 100:32–51, 2017.

[EJRHPAF19] José Estrada-Jiménez, Ana Rodríguez-Hoyos, Javier Parra-Arnau, and Jordi Forné. Measuring online tracking and privacy risks on ecuadorian websites. In *2019 IEEE Fourth Ecuador Technical Chapters Meeting (ETCM)*, pages 1–6. IEEE, 2019.

[EN16] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 1388–1401, 2016.

[ERE+15] 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.

[Eur15] Special Eurobarometer. Special eurobarometer 431: Data protection, 2015.

[Ext19] ExtUp. Cookieswap. Available at https://addons.mozilla.org/en-GB/firefox/addon/cookieswap-q/ (last accessed on 28 Feb 2020), 2019.

[Eye15] Eyeo_GmbH. Adblock plus. https://github.com/adblockplus, 2015.

[FDe13] FDev. Swap my cookie. Available at https://chrome.google.com/webstore/detail/swap-my-cookies/dffhipnliikkblkhpjapbecpmoilcama?hl=en (last accessed on 28 Feb 2020), 2013.

[FHA20] Alisa Frik, Amelia Haviland, and Alessandro Acquisti. The impact of ad-blockers on product search and purchase behavior: A lab experiment. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 163–179, 2020.

[FHUM14a] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. Anatomy of the third-party web tracking ecosystem. *arXiv preprint arXiv:1409.1066*, 2014.

[FHUM14b] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. The rise of panopticons: Examining region-specific third-party web tracking. In *International Workshop on Traffic Monitoring and Analysis*, pages 104–114. Springer, 2014.

[FVGJ19] Gertjan Franken, Tom Van Goethem, and Wouter Joosen. Exposing cookie policy flaws through an extensive evaluation of browsers and their extensions. *IEEE Security & Privacy*, 17(4):25–34, 2019.

[GCL20] Colin M Gray, Shruthi Sai Chivukula, and Ahreum Lee. What kind of work do" asshole designers" create? describing properties of ethical concern on reddit. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 61–73, 2020.

[gdp] Data protection in the eu. Available at https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en.

[GFdAS21] Danny S Guamán, Xavier Ferrer, Jose M del Alamo, and Jose Such. Automating the gdpr compliance assessment for cross-border personal data transfers in android applications. *arXiv preprint arXiv:2103.07297*, 2021.

[GFLC17a] Arthur Gervais, Alexandros Filios, Vincent Lenders, and Srdjan Capkun. Quantifying web adblocker privacy. In *European Symposium on Research in Computer Security*, pages 21–42. Springer, 2017.

[GFLC17b] Arthur Gervais, Alexandros Filios, Vincent Lenders, and Srdjan Capkun. Quantifying web adblocker privacy. In *European Symposium on Research in Computer Security*, pages 21–42. Springer, 2017.

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

[glo19] Chaffey, dave. Available at https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ (last accessed on 25 June 2021), 2019.

[GM94] John Giannandrea and Lou Montulli. Persistent client state: Http cookies, 1994.

[Goo21] Google Chrome. What are extensions? https://developer.chrome.com/docs/extensions/mv3/overview/, 2021. Online; accessed 19 Jun 2021.

[GP20] Rohit Gupta and Rohit Panda. Block the blocker: Studying the effects of anti ad-blocking. *arXiv preprint arXiv:2001.09434*, 2020.

[GRMFS13a] Richard Gomer, Eduarda Mendes Rodrigues, Natasa Milic-Frayling, and MC Schraefel. Network analysis of third party tracking: User exposure to tracking cookies through search. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 549–556. IEEE, 2013.

[GRMFs13b] Richard Gomer, Eduarda Mendes Rodrigues, Natasa Milic-Frayling, and mc schraefel. Network analysis of third party tracking: User exposure to tracking cookies through search. In *Proc. WI - IAT-Volume 01*. IEEE Computer Society, 2013.

[HdTS20] Xuehui Hu, Guillermo Suarez de Tangil, and Nishanth Sastry. Multi-country study of third party trackers from real browser histories. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 70–86. IEEE, 2020.

[HFL+18] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 531–548, 2018.

[HH19] Jeanette Herrle and Jesse Hirsh. The peril and potential of the gdpr. *Centre for International Governance Innovation*, 9, 2019.

[Hil16] Raymond Hill. ublock origin. https://github.com/gorhill/uBlock, 2016.

[HL17] Mike Hintze and Gary LaFever. Meeting upcoming gdpr requirements while maximizing the full value of data analytics. 2017.

[HL21] Niklas Hofstad and Anton Lundqvist. Cookies, cookies everywhere!: A qualitative interview study about how internet users interact with cookie consent notices, 2021.

[HLM18] Molly Hanson, Patrick Lawler, and Sam Macbeth. The tracker tax: the impact of third-party trackers on website speed in the united states. Technical report, Technical report, 2018. Available at: https://www. ghostery. com/wp-content..., 2018.

[Hof19] Chris Hoffman. Rip "do not track," the privacy standard everyone ignored. Available at https://www.howtogeek.com/fyi/rip-do-not-track-the-privacy-standard-everyone-ignored/ (last accessed on 25 June 2021), 2019.

[HP06a] KL Hui and I Png. Economics of privacy. t. hendershott, ed., handbooks in information systems, vol. 1, 2006.

[HP06b] KL Hui and I Png. Economics of privacy. t. hendershott, ed., handbooks in information systems, vol. 1, 2006.

[HP19] Thomas Hardjono and Alex Pentland. Data cooperatives: Towards a foundation for decentralized personal data management. *arXiv preprint arXiv:1905.08819*, 2019.

[HPW$^+$20] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. " it's a scavenger hunt": Usability of websites' opt-out and data deletion choices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[HS19] Xuehui Hu and Nishanth Sastry. Characterising third party cookie usage in the eu after gdpr. In *Proceedings of the 10th ACM Conference on Web Science*, pages 137–141, 2019.

[HS20] Xuehui Hu and Nishanth Sastry. What a tangled web we weave: Understanding the interconnectedness of the third party cookie ecosystem. In *12th ACM Conference on Web Science*, pages 76–85, 2020.

[HSM21] Xuehui Hu, Nishanth Sastry, and Mainack Mondal. Cccc: Corralling cookies into categories with cookiemonster. In *13th ACM Web Science Conference 2021*, pages 234–242, 2021.

[HWB20] 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.

[HZJ+19] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. An empirical analysis of data deletion and opt-out choices on 150 websites. In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*, pages 387–406, 2019.

[ICC12] ICC. Icc uk cookie guide-eu cookie law. Available at https://www.cookielaw.org/media/1096/icc_uk_cookiesguide_revnov.pdf (last accessed on 25 June 2021), 2012.

[ICO18] ICO. What is valid consent? Available at https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/consent/what-is-valid-consent/, 2018.

[ISPL18a] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. Tracing cross border web tracking. In *Proceedings of the Internet Measurement Conference 2018*, pages 329–342, 2018.

[ISPL18b] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. Tracing cross border web tracking. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, pages 329–342, New York, NY, USA, 2018. ACM.

[ISQ17] Umar Iqbal, Zubair Shafiq, and Zhiyun Qian. The ad wars: retrospective measurement and analysis of anti-adblock filter lists. In *Proceedings of the 2017 Internet Measurement Conference*, pages 171–183, 2017.

[ISZ+20] Umar Iqbal, Peter Snyder, Shitong Zhu, Benjamin Livshits, Zhiyun Qian, and Zubair Shafiq. Adgraph: A graph-based approach to ad and tracker blocking. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 763–776. IEEE, 2020.

[JG18] Ankit Kumar Jain and Brij B Gupta. Towards detection of phishing websites on client-side using machine learning based approach. *Telecommunication Systems*, 68(4):687–700, 2018.

[JG19] Ankit Kumar Jain and Brij B Gupta. A machine learning based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing*, 10(5):2015–2028, 2019.

[JSS+20] Timo Jakobi, Gunnar Stevens, Anna-Magdalena Seufert, Max Becker, and Max Von Grafenstein. Web tracking under the new data protection law: design potentials at the intersection of jurisprudence and hci. *i-com*, 19(1):31–45, 2020.

[Jun18] Nella Junge. Between privacy protection and data progression - the gdpr in the context of people analytics. *MaRBLe*, 4, 10 2018.

[KAH+20] Amir Hossein Kargaran, Mohammad Sadegh Akhondzadeh, Mohammad Reza Heidarpour, Mohammad Hossein Manshaei, Kave Salamatian, and Masoud Nejad Sattary. On detecting hidden third-party web trackers with a wide dependency chain graph: A representation learning approach. *arXiv preprint arXiv:2004.14826*, 2020.

[KAH+21] Amir Hossein Kargaran, Mohammad Sadegh Akhondzadeh, Mohammad Reza Heidarpour, Mohammad Hossein Manshaei, Kave Salamatian, and Masoud Nejad Sattary. Wide-adgraph: Detecting ad trackers with a wide dependency chain graph. In *13th ACM Web Science Conference 2021*, pages 253–261, 2021.

[Kes05] DAVID Kesmodel. When the cookies crumble. *International Business*, 9:9, 2005.

[Kla21] Daniel Kladnik. I don't care about cookies 3.3.1. https://www.i-dont-care-about-cookies.eu/, 2021.

[KNW11] Balachander Krishnamurthy, Konstantin Naryshkin, and Craig Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Proceedings of the Web*, volume 2, pages 1–10, 2011.

[Koo14] Bert-Jaap Koops. The trouble with european data protection law. *International Data Privacy Law*, 4(4):250–261, 2014.

[Kov12] Gary Kovacs. Tracking our online trackers. https://bit.ly/2gRomeS, Feb 2012. Accessed on 2018-05-12.

[Kri] Michal Krištofik. Investigation on the presence of third-parties on eu websites before and after gdpr.

[KS21] Georgios Kampanos and Siamak F Shahandashti. Accept all: The landscape of cookie banners in greece and the uk. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 213–227. Springer, 2021.

[KSA$^+$19] Florian Keusch, Bella Struminskaya, Christopher Antoun, Mick P Couper, and Frauke Kreuter. Willingness to participate in passive mobile data collection. *Public opinion quarterly*, 83(S1):210–235, 2019.

[KW09] 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.

[KWK+18] I Luk Kim, Weihang Wang, Yonghwi Kwon, Yunhui Zheng, Yousra Aafer, Weijie Meng, and Xiangyu Zhang. Adbudgetkiller: Online advertising budget draining attack. In *Proceedings of the 2018 World Wide Web Conference*, pages 297–307, 2018.

[LCJ+20] Marco Lippi, Giuseppe Contissa, Agnieszka Jablonowska, Francesca Lagioia, Hans-Wolfgang Micklitz, Przemyslaw Palka, Giovanni Sartor, and Paolo Torroni. The force awakens: Artificial intelligence for consumer law. *Journal of Artificial Intelligence Research*, 67:169–190, 2020.

[LDL+14] Mathias Lécuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. Xray: Enhancing the web's transparency with differential correlation. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 49–64, 2014.

[LGN18] Timothy Libert, Lucas Graves, and Rasmus Kleis Nielsen. Changes in third-party content on european news websites after gdpr. 2018.

[LHFE15] Tai-Ching Li, Huy Hang, Michalis Faloutsos, and Petros Efstathopoulos. Trackadvisor: Taking back browsing privacy from third-party trackers. In *International Conference on Passive and Active Network Measurement*, pages 277–289. Springer, 2015.

[Lib15] Timothy Libert. Exposing the hidden web: An analysis of third-party http requests on 1 million websites. *International Journal of Communication October*, 2015.

[Lib18] Timothy Libert. An automated approach to auditing disclosure of third-party data collection in website privacy policies. In *Proceedings of the 2018 World Wide Web Conference*, pages 207–216, 2018.

[Lin18] LinkedIn. Cookie table. https://www.linkedin.com/legal/cookie-table, May 2018. Accessed on 2018-06-01.

[LKHF20] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. The privacy policy landscape after the gdpr. *Proceedings on Privacy Enhancing Technologies*, 1:47–64, 2020.

[LN18] TL Libert and Rasmus Kleis Nielsen. Third-party web content on eu news sites: Potential challenges and paths to privacy improvement. 2018.

[Lon18] King's College London. Research ethics. https://bit.ly/2vjh8ah, Mar 2018. Accessed on 2018-03-22.

[Low] Peter Lowe. pgl.yoyo.org blocklist. Available at http://pgl.yoyo.org/as/serverlist.php?hostformat=nohtml&showintro=1&mimetype=plaintext.

[LR15] Timothy Libert and Maria Repnikova. Google is returning to china? it never really left. Available at https://www.theguardian.com/technology/2015/sep/21/google-is-returning-to-china-it-never-really-left (last accessed on 25 June 2021), 2015.

[LSW+13a] Bin Liu, Anmol Sheth, Udi Weinsberg, Jaideep Chandrashekar, and Ramesh Govindan. Adreveal: Improving transparency into online targeted advertising. In *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*, pages 1–7, 2013.

[LSW+13b] Bin Liu, Anmol Sheth, Udi Weinsberg, Jaideep Chandrashekar, and Ramesh Govindan. Adreveal: improving transparency into online targeted advertising. In *12th Proc. HotNets*, page 12. ACM, 2013.

[LUW+13] Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. What matters to users? factors that affect users' willingness to share information with online advertisers. In *Proceedings of the ninth symposium on usable privacy and security*, pages 1–12, 2013.

[LWED20] Ze Shi Li, Colin Werner, Neil Ernst, and Daniela Damian. Gdpr compliance in the context of continuous integration. *arXiv preprint arXiv:2002.06830*, 2020.

[LZC16] Sakchan Luangmaneerote, Ed Zaluska, and Leslie Carr. Survey of existing fingerprint countermeasures. In *2016 International Conference on Information Society (i-Society)*, pages 137–141. IEEE, 2016.

[Mad20] Gerald Madlmayr. Minimal consent - browsing made seamless again. https://www.minimal-consent.com/, 2020.

[MAF$^+$19] Arunesh Mathur, Gunes Acar, Michael J Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32, 2019.

[Mal] MalwareTips. Malware removal guides. Available at https://malwaretips.com/forums/malware-removal-guides.11.

[Mal17] Gabriel Maldoff. The risk-based approach in the gdpr: interpretation and implications. *IAPP https://iapp.org/media/pdf/resource_center/GDPR_Study_Maldoff. pdf. Accessed*, 12, 2017.

[Mar06] Dave Martorana. Multifirefox. Available at https://davemartorana.com/multifirefox/ (last accessed on 28 Feb 2020), 2006.

[MBMC18] Daniele Moro, Filippo Benati, Michele Mangili, and Antonio Capone. Catching free-riders: in-network adblock detection with machine learning techniques. In *2018 IEEE 23rd International Workshop on Computer*

Aided Modeling and Design of Communication Links and Networks (CAMAD), pages 1–6. IEEE, 2018.

[MBS20] Célestin Matte, Nataliia Bielova, and Cristiana Santos. Do cookie banners respect my choice?: Measuring legal compliance of banners from iab europe's transparency and consent framework. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 791–809. IEEE, 2020.

[MGF19] Johan Mazel, Richard Garnier, and Kensuke Fukuda. A comparison of web privacy protection techniques. *Computer Communications*, 144:162–174, 2019.

[MHB^{+}17] Georg Merzdovnik, Markus Huber, Damjan Buhov, Nick Nikiforakis, Sebastian Neuner, Martin Schmiedecker, and Edgar Weippl. Block me if you can: A large-scale study of tracker-blocking tools. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 319–333. IEEE, 2017.

[MLXL16] Wei Meng, Byoungyoung Lee, Xinyu Xing, and Wenke Lee. Trackmeornot: Enabling flexible control on web tracking. In *Proceedings of the 25th International Conference on World Wide Web*, pages 99–109, 2016.

[MM12] Jonathan R Mayer and John C Mitchell. Third-party web tracking: Policy and technology. In *2012 IEEE symposium on security and privacy*, pages 413–427. IEEE, 2012.

[Moz12] Mozilla. Firefox lightbeam. Available at https://github.com/mozilla/lightbeam-we, 2012.

[Moz13] Mozilla. Collusion. https://wiki.mozilla.org/Collusion, Jun 2013. Accessed on 2018-07-11.

[Moz15] Mozilla. Multi-account containers. Available at https://support.mozilla.org/en-US/kb/containers#w_what-are-containers (last accessed on 24 June 2021), 2015.

[Moz17] Mozilla. Firefox multi-account containers. Available at {https://addons.mozilla.org/en-GB/firefox/addon/multi-account-containers/versions/s}(lastaccessedon24June2021), 2017.

[Moz18a] Mozilla. Facebook container - prevent facebook from tracking you on other websites. Available at https://support.mozilla.org/en-US/kb/facebook-container-prevent-facebook-tracking (last accessed on 25 June 2021), 2018.

[Moz18b] Mozilla. Facebook container extension: Take control of how you're being tracked. Available at https://blog.mozilla.org/firefox/facebook-container-extension/ (last accessed on 24 June 2021), 2018.

[Moz18c] Mozilla. Webrequest api. https://mzl.la/2ZcaqjQ, Jul 2018. Accessed on 2018-07-22.

[Moz21] Mozilla. Contextual identities. Available at https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/API/contextualIdentities (last accessed on 24 June 2021), 2021.

[MPS+13] Delfina Malandrino, Andrea Petta, Vittorio Scarano, Luigi Serra, Raffaele Spinelli, and Balachander Krishnamurthy. Privacy awareness about information leakage: Who knows what about me? In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pages 279–284, 2013.

[MQS17] Muhammad Haris Mughees, Zhiyun Qian, and Zubair Shafiq. Detecting anti ad-blockers in the wild. *Proceedings on Privacy Enhancing Technologies*, 2017(3):130–146, 2017.

[MSH+19] Nicole A Maher, Joeky T Senders, Alexander FC Hulsbergen, Nayan Lamba, Michael Parker, Jukka-Pekka Onnela, Annelien L Bredenoord, Timothy R Smith, and Marike LD Broekman. Passive data collection and use in healthcare: A systematic review of ethical issues. *International journal of medical informatics*, 129:242–247, 2019.

[MTM15a] H. Metwalley, S. Traverso, and M. Mellia. Unsupervised detection of web trackers. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2015.

[MTM+15b] Hassan Metwalley, Stefano Traverso, Marco Mellia, Stanislav Miskovic, and Mario Baldi. The online tracking horde: a view from passive measurements. In *International Workshop on Traffic Monitoring and Analysis*, pages 111–125. Springer, 2015.

[MTM+15c] Hassan Metwalley, Stefano Traverso, Marco Mellia, Stanislav Miskovic, and Mario Baldi. The online tracking horde: a view from passive measurements. In *International Workshop on Traffic Monitoring and Analysis*, pages 111–125. Springer, 2015.

[MTM16] Hassan Metwalley, Stefano Traverso, and Marco Mellia. Using passive measurements to demystify online trackers. *Computer*, 49(3):50–55, 2016.

[Net02] Netscape. Persistent client state http cookies. Available at https://bit.ly/3qY55Ks (last accessed on 25 June 2021), 2002.

[NHH07] Patricia A Norberg, Daniel R Horne, and David A Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of consumer affairs*, 41(1):100–126, 2007.

[NLV+20] 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.

[One18] OneTrust. Onetrust launches universal consent management and preference management. https://bit.ly/2V674Qq, March 2018. Accessed on 2019-02-10.

[One19] OneTrust. Onetrust preferencechoice's cookie auto-blocking technology. https://bit.ly/2O4HnO6, 2019.

[One20] OneTrust. Cookiepedia. https://cookiepedia.co.uk/, 2020.

[Pap18] Panagiotis Papadopoulos. Analyzing the impact of digitaladvertising on user privacy, 2018.

[Par00] Canadian Parliament. Personal information protection and electronic documents act. *Consolidated Acts, SC 2000, c*, 5:13, 2000.

[PCT06] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*, pages 1–es, 2006.

[PKM19] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference*, pages 1432–1442, 2019.

[PO16a] Marc Pelteret and Jacques Ophoff. A review of information privacy and its importance to consumers and organizations. *Informing Science*, 19:277–301, 2016.

[PO16b] Marc Pelteret and Jacques Ophoff. A review of information privacy and its importance to consumers and organizations. *Informing Science: The International Journal of an Emerging Transdiscipline*, 19:277–301, 2016.

[Por18] GDPR Portal. Eu gdpr information portal. https://www.eugdpr.org/, May 2018. Accessed on 2018-06-11.

[PPA+20] Phu H Phung, Huu-Danh Pham, Jack Armentrout, Panchakshari N Hiremath, and Quang Tran-Minh. A user-oriented approach and tool for security and privacy protection on the web. *SN Computer Science*, 1(4):1–16, 2020.

[PPAB16] Sören Preibusch, Thomas Peetz, Gunes Acar, and Bettina Berendt. Shopping for privacy: Purchase details leaked to paypal. *Electronic Commerce Research and Applications*, 15:52–64, 2016.

[PPKM21] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P Markatos. User tracking in the post-cookie era: How websites bypass gdpr consent to track users. In *Proceedings of the Web Conference 2021*, pages 2130–2141, 2021.

[Pre18] Presidência da República. Lei geral de proteção de dados pessoais (lgpd), 2018.

[Pri20] PrivacyCloud. Consent manager. https://chrome.google.com/webstore/detail/consent-manager/gpkoajillfmlpnglbagpplnphadbfalh?hl=en, 2020.

[PRMF16] Silvia Puglisi, David Rebollo-Monedero, and Jordi Forné. On web user tracking: How third-party http requests track users' browsing patterns for personalised advertising. In *2016 Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, pages 1–6. IEEE, 2016.

[PSL19] Panagiotis Papadopoulos, Peter Snyder, and Benjamin Livshits. Another brick in the paywall: The popularity and privacy implications of paywalls. *CoRR, abs/1903.01406*, 2019.

[PXQ+19] Peng Peng, Chao Xu, Luke Quinn, Hang Hu, Bimal Viswanath, and Gang Wang. What happens after you leak your password: Understanding credential sharing on phishing sites. 2019.

[Qua21] Quantcast. Quantcast choice - user guide. Available at https://help.quantcast.com/hc/en-us/articles/360052725133-Quantcast-Choice-User-Guide, 2021.

[Red18] Reddit. Ad personalization. https://www.reddit.com/personalization, May 2018. Accessed on 2018-05-30.

[Red19] Elissa M Redmiles. " should i worry?" a cross-cultural examination of account security incident response. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 920–934. IEEE, 2019.

[RF09] Guillaume Roels and Kristin Fridgeirsdottir. Dynamic revenue management for online display advertising. *Journal of Revenue and Pricing Management*, 8(5):452–466, 2009.

[RKDK17] Philip Raschke, Axel Küpper, Olha Drozd, and Sabrina Kirrane. Designing a gdpr-compliant and usable privacy dashboard. In *IFIP International Summer School on Privacy and Identity Management*, pages 221–236. Springer, 2017.

[RKW12] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*, pages 155–168, 2012.

[RNVR+18] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018.

[RSBL] KV Rosni, Manish Shukla, Vijayanand Banahatti, and Sachin Lodha. Consent recommender system: A case study on linkedin settings.

[RSS+16] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. Expecting the unexpected: Understanding mismatched privacy expectations online. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 77–96, 2016.

[Sal15] Johnny M. Saldana. *The coding manual for qualitative researcher*. SAGE, 3 edition, 2015.

[SBK16] Sophie Stalla-Bourdillon and Alison Knight. Anonymous data v. personal data-false debate: An eu perspective on anonymization, pseudonymization and personal data. *Wis. Int'l LJ*, 34:284, 2016.

[Sch12] Daniel Schneider. Data protection in germany. In *Trust in Biobanking*, pages 169–187. Springer, 2012.

[SCL18] Yong Shi, Gong Chen, and Juntao Li. Malicious domain name detection based on extreme machine learning. *Neural Processing Letters*, 48(3):1347–1357, 2018.

[SCM+10] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. Flash cookies and privacy. In *2010 AAAI Spring Symposium Series*, 2010.

[Sel21] Selenium. Selenium webdriver, 2021.

[Sen19] Senzing. Finding the missing link in gdpr compliance. https://bit.ly/2IGzUjA, January 2019. Accessed on 2019-02-08.

[SIIK19] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. Clash of the trackers: Measuring the evolution of the online tracking ecosystem. *arXiv preprint arXiv:1907.12860*, 2019.

[SK16a] Sebastian Schelter and Jérôme Kunegis. Tracking the trackers: A large-scale analysis of embedded web trackers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, 2016.

[SK16b] Sebastian Schelter and Jérôme Kunegis. Tracking the trackers: A large-scale analysis of embedded web trackers. In *ICWSM*, pages 679–682, 2016.

[SK18] Sebastian Schelter and Jérôme Kunegis. On the ubiquity of web tracking: Insights from a billion-page web crawl. *Journal of Web Science*, 4:53–66, 2018.

[SK19] Jannick Sørensen and Sokol Kosta. Before and after gdpr: The changes in third party presence at public and private european websites. In *The World Wide Web Conference*, pages 1590–1600, 2019.

[SM19a] Takahito Sakamoto and Masahiro Matsunaga. After gdpr, still tracking or not? understanding opt-out states for online behavioral advertising. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 92–99. IEEE, 2019.

[SM19b] Nayanamana Samarasinghe and Mohammad Mannan. Towards a global perspective on web tracking. *Computers & Security*, 87:101569, 2019.

[SNT$^+$21] Cristiana Santos, Midas Nouwens, Michael Toth, Nataliia Bielova, and Vincent Roca. Consent management platforms under the gdpr: processors and/or controllers? In *Annual Privacy Forum*, pages 47–69. Springer, 2021.

[SPK16] Suphannee Sivakorn, Iasonas Polakis, and Angelos D Keromytis. The cracked cookie jar: Http cookie hijacking and the exposure of private information. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 724–742. IEEE, 2016.

[SRDK+19a] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. Can i opt out yet? gdpr and the global illusion of cookie control. In *Proceedings of the 2019 ACM Asia conference on computer and communications security*, pages 340–351, 2019.

[SRDK+19b] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. Can i opt out yet?: Gdpr and the global illusion of cookie control. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, pages 340–351. ACM, 2019.

[SRSA19] Ahmedur Rahman Shovon, Shanto Roy, Arnab Kumar Shil, and Tanjila Atik. Gdpr compliance: Implementation use cases for user data privacy in news media industry. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6. IEEE, 2019.

[Sta18] State of California Department of Justice. California consumer privacy act (ccpa), 2018.

[Sta21a] StatCounter Global Stats. Desktop browser market share worldwide | statcounter global stats. https://gs.statcounter.com/browser-market-share/desktop/worldwide/#monthly-202106-202106-bar, 2021. Online; accessed 18 Aug. 2021.

[Sta21b] StatCounter Global Stats. Search engine market share worldwide. Available at https://gs.statcounter.com/search-engine-market-share/all, 2021.

[SVdB20] Jannick Kirk Sørensen and Hilde Van den Bulck. Public service media online, advertising and the third-party user data business: A trade versus trust dilemma? *Convergence*, 26(2):421–447, 2020.

[SVM13] L Stevenson, R Vareia, and JA Moren. Sustaining colombians through the entrepreneurial pipeline–a policy challenge for colombia? *Global Entrepreneurship Monitor Report, Ottawa, International Development Research Centre*, 2013.

[SWS15] Marian Harbach Susanne Weber and Matthew Smith. Participatory Design for Security-Related User Interfaces. In *Proceedings of the Workshop on Usable Security (USEC'15)*, San Diego, CA, February 2015.

[Tan16] Colin Tankard. What the gdpr means for businesses. *Network Security*, 2016(6):5–8, 2016.

[TGM15] Wiebke Thode, Joachim Griesbaum, and Thomas Mandl. " i would have never allowed it": User perception of third-party tracking and implications for display advertising. In *ISI*, pages 445–456, 2015.

[Tho20] Reuben Thomas. Enchant. Available at https://abiword.github.io/enchant/ (last accessed on 25 June 2021), 2020.

[TJH+18] Ke Tian, Steve TK Jan, Hang Hu, Danfeng Yao, and Gang Wang. Needle in a haystack: Tracking down elite phishing domains in the wild. In *Proceedings of the Internet Measurement Conference 2018*, pages 429–442, 2018.

[Too12] Toolness. Collusion. Available at http://www.toolness.com/wp/2011/07/collusion/, 2012.

[TPRM18] Christina Tikkinen-Piri, Anna Rohunen, and Jouni Markkula. Eu general data protection regulation: Changes and implications for personal data collecting companies. *Computer Law & Security Review*, 34(1):134–153, 2018.

[Tru] TrustedSource. Customer url ticketing system. Available at www.trustedsource.org/en/feedback/url (last accessed on 25 June 2021).

[Tru18] TrustArc. Trustarc launches gdpr validation, empowering companies to demonstrate gdpr compliance status. https://bit.ly/2PgBaet, April 2018. Accessed on 2019-02-10.

[Twi18] Twitter. Ad personalization. https://twitter.com/settings/personalization, May 2018. Accessed on 2018-05-30.

[UDF^{+}19] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (un) informed consent: Studying gdpr consent notices in the field. In *Proceedings of the 2019 acm sigsac conference on computer and communications security*, pages 973–990, 2019.

[UDHP20] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Beyond the front page: Measuring third party dynamics in the field. *arXiv preprint arXiv:2001.10248*, 2020.

[Uni19] Aarhus University. Consent-o-matic. https://github.com/cavi-au/Consent-O-Matic, 2019.

[UTD^{+}18] 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.

[UTD^{+}20] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Measuring the impact of the gdpr on data sharing in ad networks. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, pages 222–235, 2020.

[VEAWN19] 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). IEEE*, 2019.

[VEMR$^+$18] Natalija Vlajic, Marmara El Masri, Gianluigi M Riva, Marguerite Barry, and Derek Doran. Online tracking of kids and teens by means of invisible images: Coppa vs. gdpr. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, pages 96–103, 2018.

[VMK17] Tanvi Vyas, Andrea Marchesini, and Christoph Kerschbaumer. Extending the same origin policy with origin attributes. In *ICISSP*, pages 464–473, 2017.

[Wal20] Ari Ezra Waldman. Cognitive biases, dark patterns, and the 'privacy paradox'. *Current opinion in psychology*, 31:105–109, 2020.

[Who] WhoTracks.me. Trackers rank. Available at https://whotracks.me/trackers.html.

[WLZW15] Qianru Wu, Qixu Liu, Yuqing Zhang, and Guanxing Wen. Trackerdetector: A system to detect third-party trackers through machine learning. *Computer Networks*, 91:164–173, 2015.

[WS19] Angela G Winegar and Cass R Sunstein. How much is data privacy worth? a preliminary investigation. *Journal of Consumer Policy*, 42(3):425–440, 2019.

[WU16] Craig E Wills and Doruk C Uzunoglu. What ad blockers are (and are not) doing. In *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, pages 72–77. IEEE, 2016.

[Xue20] Yiqun Xue. *Make a More Meaningful Interaction: Exploring the Framework of Cookie Notice*. PhD thesis, Doctoral dissertation, Waseda University, 2020.

[Zar16a] Tal Z Zarsky. Incompatible: the gdpr in the age of big data. *Seton Hall L. Rev.*, 47:995, 2016.

[Zar16b] Tal Z Zarsky. Incompatible: The gdpr in the age of big data. *Seton Hall L. Rev.*, 47:995, 2016.

[Zar17] Tal Z. Zarsky. Incompatible: The gdpr in the age of big data. *The Seton Hall Law Review*, 47:2, 2017.

[ZB20] Razieh Nokhbeh Zaeem and K Suzanne Barber. The effect of the gdpr on privacy policies: recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)*, 12(1):1–20, 2020.

[ZCBZ19] Hong Zhao, Zhaobin Chang, Guangbin Bao, and Xiangyan Zeng. Malicious domain names detection algorithm based on n-gram. *Journal of Computer Networks and Communications*, 2019, 2019.

[Zeu] Zeus. Zeus tracker. Available at https://zeustracker.abuse.ch/blocklist.php.

[Zhe17] Lili Zheng. Does online perceived risk depend on culture? individualistic versus collectivistic culture. *Journal of Decision Systems*, 26(3):256–274, 2017.

[ZHQ+18] Shitong Zhu, Xunchao Hu, Zhiyun Qian, Zubair Shafiq, and Heng Yin. Measuring and disrupting anti-adblockers using differential execution analysis. In *The Network and Distributed System Security Symposium (NDSS)*, 2018.