Advanced Methods to Audit Online Web Services.

by

Pelayo Vallina Rodríguez

A dissertation submitted by in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Telematic Engineering

Universidad Carlos III de Madrid

Advisor(s):

Prof. Dr. Antonio Fernández Anta
Prof. Dr. Rubén Cuevas Rumín

Tutor:

Prof. Dr. Antonio Fernández Anta

November 2022

Advanced Methods to Audit Online Web Services.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Prepared by:
Pelayo Vallina Rodríguez, IMDEA Networks Institute / Universidad Carlos III de Madrid

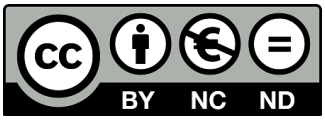
Under the advice of:
Antonio Fernández Anta, IMDEA Networks Institute
Rubén Cuevas Rumín, Universidad Carlos III de Madrid

Departamento de Ingeniería Telemática, Universidad Carlos III de Madrid

---

Contact: pelayo.vallina@imdea.org

This work has been supported by IMDEA Networks Institute.



ACKNOWLEDGEMENTS

First and foremost, I have to thank my supervisors, Antonio Fernandez Anta and Rubén Cuevas Rumín for their help, advice, patience, and support. Thanks for giving me the great opportunity of doing a PhD. with both of you. You are the primary reason why I am here today.

I am very grateful to all my collaborators, Ángel Cuevas Rumín, Antonio Nappa, Antonio Pastor, José González Cabañas, Narseo Vallina, Oliver Hohlfeld, Patricia Callejo, Roberto Gonzalez, Sergio Pastrana, and victor Le Pochat. Thanks to all IMDEA Networks and UC3M friends, for all the trips and all the good moments together, Alvaro, Julien, Segun, Constantine, Aniketh, Srdjan, Guillermo, Roberto, Andres, Hany, Roderick, Jorge, Victor, Borja, José, Antonio, Carlos, Marco, and Patricia. Sorry to all those I did not mention.

I'd like to thank Gareth Tyson and Ignacio Castro for giving me the honor of doing an internship with you at the Queen Mary University of London, and for having such a great time there. I would like to show my gratitude also to all the professors and students who made me feel at home during my period at QMUL. Thanks to Alberto Pozanco Lancho for hosting me in London, it was like coming back 10 years ago when we lived together. This time there was not any project involved together, beyond going out.

Finally, I would like to show my immense gratitude to my parents, brother, Bea and of course, Gersán and Deva for bringing happiness and joy to all of us. Thanks to my girlfriend, Arancha, for your support, love, and patience in times of difficulty. I'm eternally grateful, as without you it would not have been possible.

ABSTRACT

Online web services have grown dramatically in size and diversity in the last years, becoming essential components of our daily life and allowing us to conduct elementary tasks like working, getting informed, or keeping in contact with relatives and friends.

However, all the changes and evolution experimented on by the online web services had not have been possible without implementing a profitable economic model that sustains it. Despite a suitable percentage of these services being fee-based, they represent a lucrative business that generates billions of dollars, allowing the creation of some of the biggest companies in the world in terms of market capitalization, like Alphabet Inc. or Meta Inc. (Previously known as Facebook Inc.). Being costless and lucrative is possible due to an advertising-based monetization model, which consists of delivering ads to the users in exchange for their services (*e.g.,* Facebook or YouTube). Although online advertising dates back to the middle of the 90s, its popularity has experienced an increase among brands and advertising agencies in the last decade, mainly due to its capacity to reach precise audiences at a low cost.

Converting online web services into advertising walls is a double-edged sword for the users. The capacity offered by online advertising to segment their audiences requires a massive collection of personal data from the users, including their web browsing histories or even more invasive data such as age, gender, or location to infer the online profile of the users. This data collection is possible due to implementing a complex tracking ecosystem by online advertising companies from which multiple stakeholders collect, process, and exchange information. The many privacy cases of abuse inflicted by this industry motivated the implementation of new regulatory efforts to protect consumers' privacy in the last years. Some notable examples are the General Data Protection Regulation (GDPR) [1] in the European Union or the California Consumer Privacy Act Regulations (CCPA) [2] in California, USA. Further, these privacy regulations typically contain specific provisions and strict requirements for websites that provide sensitive material to end users, including sexual, religious, and health services.

Implementing new regulatory frameworks, alongside the growth of online web services, forces an endless evolution of current techniques to study and audit online web services. Furthermore, there is a need to emphasize the online advertising ecosystem, as it represents the primary economic support of a high percentage of web services. Also, the activities and abuses conducted by this ecosystem drove the implementation of current privacy regulations to control the use and collection of personal data.

This dissertation falls within the topics of Internet measurements, tackling the need for new measurement techniques and methodological approaches to audit and study online web services. These efforts want to increase the limited

knowledge about web subsystems offering sensitive material, including their regulatory compliance regarding current privacy regulations. Also, this dissertation tackles the need to study and measure how big ad tech companies create and use the online profiles of their users to distribute tailored ads. Furthermore, the work presented in this dissertation raises the need for a more in-depth understanding of fundamental tools for conducting Internet measurement works, including their limitations and suitability for academic research. Specifically, this dissertation presents three main contributions:

The first one corresponds with implementing a novel methodology to audit sensitive web services' privacy, transparency, and regulatory compliance. We validate our method by looking at pornographic websites concerning the GDPR in the European Union. We focus our analysis on such types of websites for two main reasons: (*i*) the GDPR establishes specific provisions and strict requirements on sensitive websites, including pornographic ones. (*ii*) big ad tech companies set strict constraints for porn-related publishers. As a result, it opened new market opportunities for other actors who have specialized in advertising and tracking technologies for adult sites, creating a semi-decoupled ecosystem from the rest of the web. We perform a holistic analysis of over 6,843 pornographic websites, finding a prevalent absence of regulatory compliance and very extended use of tracking techniques, including advanced ones such as fingerprinting. These results stress the importance of studying the World Wide Web subsets that have not been scrutinized by regulators, policymakers, and the research community in depth.

Second, we empirically and comprehensively analyze 13 domain classification services to study their labeling strategy and performance. These services have multiple applications, from business applications such as online advertising to academic research works to conduct category-dependent measurements or to identify the purpose of a website or online service. We study each domain classification service's methodologies, scalability limitations, label constellations, and suitability for academic research studies. In some cases, their findings depend on the results provided by the domain classification services. We find that the limitations and shortcomings of each domain classification service heavily affect their suitability and applicability, both for practical solutions and academic studies.

In the third and last contribution, we implement a novel methodology with real users to study the performance and quality of the profiling and ad targeting algorithms from the two most important stakeholders in the online advertising business, Google and Meta (previously Facebook). We find that half of the categories associated with the profiles are incorrectly assigned. We also observe the presence of sensitive categories in Facebook users, posing a privacy risk and potential regulatory noncompliance.

In summary, this dissertation brings new methodologies and results to increase our limited knowledge about the web.

PUBLISHED AND SUBMITTED CONTENT

The content of this dissertation has been published in the following conferences and journals:

1. **Pelayo Vallina**, Álvaro Feal, Julien Gamba, Antonio Fernández Anta, and Narseo Vallina-Rodriguez. "Tales from the Porn: A Comprehensive Privacy Analysis of the Web Porn Ecosystem". Published in *ACM Internet Measurement Conference 2019 (ACM IMC 2019)*. https://dl.acm.org/doi/10.1145/3355369.3355583
   - This work is fully included and its content is reported in Chapter 3.
2. **Pelayo Vallina**, Álvaro Feal, Julien Gamba, Antonio Fernández Anta, and Narseo Vallina-Rodriguez. "Everybody's got something to hide: Analysis of the Online Adult Ecosystem". Published in *ACM Internet Measurement Conference 2018 (Poster at ACM IMC 2018)*. https://dspace.networks.imdea.org/handle/20.500.12761/641
   - This work is fully included and its content is reported in Chapter 3.
3. **Pelayo Vallina**, Álvaro Feal, Julien Gamba, Antonio Fernández Anta, and Narseo Vallina-Rodriguez. "This Is My Private Business! Privacy Risks on Adult Websites". Published in *IV Jornadas Nacionales de Investigación en Ciberseguridad (JNIC 2018)*. https://dspace.networks.imdea.org/handle/20.500.12761/606
   - This work is fully included and its content is reported in Chapter 3.
4. **Pelayo Vallina**, Victor Le Pochat, Álvaro Feal, Marius Paraschiv, Julien Gamba, Tim Burke , Oliver Hohlfeld, Juan E. Tapiador, and Narseo Vallina-Rodriguez. "Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification". Published in *ACM Internet Measurement Conference 2020 (ACM IMC 2020)*. https://dl.acm.org/doi/10.1145/3419394.3423660
   - This work is fully included and its content is reported in Chapter 4.
5. **Pelayo Vallina**, José González-Cabañas, Rubén Cuevas, Antonio Fernández Anta, and Ángel Cuevas. "You Don't Know Me: Auditing the performance of Google and Facebook's profiling and ad targeting algorithms". *In Preparation*.
   - This work is fully included and its content is reported in Chapter 5.

OTHER PUBLICATIONS

During my time working on this dissertation, I had the pleasure of co-authoring other publications, which are not included in this document.

1. **Pelayo Vallina**, Ignacio Castro, and Gareth Tyson. "Cashing in on Contacts: Characterizing the OnlyFans Ecosystem". In Preparation.

2. José Miguel Moreno, Sergio Pastrana, Jens Helge Reelfs, **Pelayo Vallina**, Andriy Panchenko, Georgios Smaragdakis, Oliver Hohlfeld, Narseo Vallina-Rodriguez and Juan Tapiador. "Let Me Inform You! Bypassing Russian Censorship in War Times". In Preparation.

3. José González-Cabañas, Patricia Callejo, **Pelayo Vallina**, Ángel Cuevas, Rubén Cuevas, and Antonio Fernández-Anta. "How resilient is the Open Web to the COVID-19 pandemic?". Published in *Elsevier Journal of Telematics and Informatics 2021*. https://doi.org/10.1016/j.tele.2021.101692

4. Álvaro Feal, **Pelayo Vallina**, Julien Gamba, Sergio Pastrana, Antonio Nappa, Oliver Hohlfeld, Narseo Vallina-Rodriguez, Juan Tapiador. "Blocklist Babel: On the Transparency and Dynamics of Open Source Blocklisting". Published in *IEEE Transactions on Network and Service Management 2021*. https://ieeexplore.ieee.org/document/9416274

5. Antonio Pastor, Matti Pärssinen, Patricia Callejo, **Pelayo Vallina**, Rubén Cuevas, Ángel Cuevas, Mikko Kotila and Arturo Azcorra. "Nameles: An intelligent system for Real-Time Filtering of Invalid Ad Traffic". Published in *The World Wide Web Conference (ACM WWW 2019)*. https://doi.org/10.1145/3308558.3313601

6. **Pelayo Vallina**, Antonio Fernández-Anta, Rubén Cuevas and Esteban Moro. "MyBubble: Influence of Algorithms in Users' Filter Bubbles". *Poster at OPERANDI 2018, co-located with PoPETs 2018*. https://dspace.networks.imdea.org/bitstream/handle/20.500.12761/653/operandi_poster.pdf

CONTENTS

LIST OF FIGURES

ACRONYMS

ASACP: Association of Sites Advocating Child Protection

ATS: Advertisement and Tracking Services

CCPA: California's Consumer Privacy Act

CCS: Computing Classification System

CDN: Content Delivery Networks

CSync: Cookies Synchronization

CoNEXT: Conference on emerging Networking EXperiments and Technologies

DAU: Daily Active Users

DPO: Data Protection Officer

EU: European Union

FB: Facebook

FQDN: Fully Qualified Domain Name

GDPR: General Data Protection Regulation

IAB: Internet Advertising Bureau

ID: Identifier

IETF: Internet Engineering Task Force

IMC: Internet Measurement Conference

IRB: Institutional Ethics Board

MAU: Monthly Active Users

NDSS: Network and Distributed System Security symposium

OBA: Online Behavioural Advertising

PAM: Passive and Active Measurement conference

PETS: Privacy Enhancing Technologies Symposium

PII: Personally Identifiable Information

RTA: Restricted-for-Adults

S&P: Security and Privacy

TF-IDF: Term Frequency-Inverse Document Frequency

TMA: network Traffic Measurement and Analysis conference

WWW: World Wide Web

Part I

INTRODUCTION AND BACKGROUND

1

INTRODUCTION

Online web services have appeared, evolved, and grown in size, diversity, and complexity in the last decades, becoming essential parts of our lives. Nowadays, we rely on them to interact with our friends and relatives, do our jobs, make purchases, or stay informed in real-time about relevant events that occur locally and globally. This process has become even more evident since the COVID-19 pandemic, when millions of citizens have depended on them to perform basic actions.

However, all the changes and evolution experimented by the online web services had not have been possible without implementing a profitable economic model that sustains it. Despite a suitable percentage of these services being fee-based, they represent a lucrative business that generates billions of dollars. The profits generated by these companies made them some of the biggest companies in the world in terms of market capitalization, like Alphabet Inc. or Meta Inc. (Previously known as Facebook Inc.). Furthermore, free access is possible due to implementing an advertising-based monetization model, allowing their economic viability. Online web services offer spaces in their platforms to advertisers, who buy them to place their ads. Users receive such ads when they visit and interact with the websites..

The capacity of online advertising to deliver personalized ads (segmenting the audiences) and their low cost compared to traditional mainstream advertising channels, including TV, newspapers, billboards, or radio, has raised its popularity among brands and advertising agencies. Consequently, the companies that are part of the online advertising ecosystem have experimented a continuous revenue growth since its creation. According to the Internet Advertising Bureau (IAB), [3], these companies have reached annual gains of over 15% [4].

The online advertising ecosystem has built a business model based on the collection and process of personal data, leading to many privacy cases of abuse that, together with scandals like Cambridge Analytica [5], have motivated new regulatory efforts to protect consumers' privacy. These new regulations want to give the users control of their data, imposing very clear directives on how the personal data should be stored, processed, collected (under informed consent from users), processed, and under which circumstances. Some of the most advanced efforts to protect the privacy of the users are the General Data Protection Regulation (GDPR) [1] in the European Union or the California Consumer Privacy Act Regulations (CCPA) [2] in California, USA. Further, these new regulatory frameworks impose additional requirements and restrictions on websites that provide sensitive material to end users, like sexual, religious, and health-related services.

The new regulatory frameworks, alongside the growth of online web services, force an endless evolution of current techniques to study and audit online

web services. Furthermore, there is a need to emphasize the online advertising ecosystem, as it represents the primary economic support of a high percentage of web services. Also, the activities and abuses conducted by this ecosystem drove the implementation of current privacy regulations to control the use and collection of personal data.

This dissertation falls within the topics of Internet measurements, tackling the need for new measurement techniques and methodological approaches to audit and study online web services. These efforts try to increase the limited knowledge about opaque web subsystems and how big ad tech companies create and use the online profiles of their users to distribute tailored ads. Furthermore, the work presented in this dissertation raises the need for a more in-depth understanding of fundamental tools for conducting Internet measurement works, including their limitations and suitability for academic research.

Precisely, this dissertation analyzes three aspects of the web. Firsts, we implement a methodology to study sensitive websites, including their potential lack of regulatory compliance. Then, we put into practice our approach by analyzing the pornographic web ecosystem, opening the debate on the need to study and identify web privacy problems from a macroscopic perspective, as the web contains semi-decoupled and highly sensitive subsystems. Second, we look deeply at the suitability and adequacy of domain classification services commonly used by the research community to conduct domain-dependent research studies, including those studying sensitive websites. Finally, we implement a novel methodology to audit the quality and performance of the profiles that Meta (Facebook) and Google create about the users and their ad targeting algorithms. This study also includes an analysis of the transparency tools these two companies offer to the users concerning the process of distributing tailored ads.

1.1 CONTRIBUTIONS

This dissertation provides novel methodologies and tools to audit and analyze web services and the suitability of fundamental tools for conducting Internet measurement works. This dissertation provides three main contributions:

1. **Privacy Analysis of the Web Porn Ecosystem.** Current privacy regulations, including the General Data Protection Regulation (GDPR) [1] in the European Union, aim to control user-tracking activities in websites and mobile applications. These privacy rules typically contain specific provisions and strict requirements for websites that provide sensitive material to end users, such as sexual, religious, and health services. However, little is known about users' privacy risks when visiting such websites and their regulatory compliance. Previous research works have analyzed and studied these aspects of the web as a monolithic ecosystem without considering the presence of web subsystems and their particular requirements. We present the first comprehensive and large-scale analysis of pornographic websites. We provide an exhaustive behavioral analysis of the use of tracking methods by 6,843 pornographic websites and a study of their lack of

regulatory compliance, including the absence of age-verification mechanisms to prevent minors' access to porn websites and methods to obtain informed user consent. The results indicate that, as in the regular web, tracking is prevalent across pornographic sites: 72% of the websites use third-party cookies and 5% leverage advanced user fingerprinting technologies. Yet, our analysis reveals a third-party tracking ecosystem semi-decoupled from the regular web in which various analytics and advertising services track users across and outside pornographic websites. Finally, we complete the study with a regulatory compliance analysis in the context of the EU GDPR and the newer legal requirements to implement verifiable access control mechanisms. We find that only 16% of the analyzed websites have an accessible privacy policy, and only 4% provide a cookie consent banner. The use of verifiable access control mechanisms is limited to prominent pornographic websites.

2. **An Analysis of Domain Classification Services.** Domain classification services have applications in multiple areas, including cybersecurity, content blocking, and targeted advertising. Yet, these services are often a black box in terms of their methodology for classifying domains, making it difficult to assess their strengths, aptness for specific applications, and limitations. In Chapter 4, we perform a large-scale analysis of 13 popular domain classification services on more than 4.4 million hostnames. We empirically explore their methodologies, scalability limitations, label constellations, and their suitability for academic research, as well as other practical applications such as content filtering. We find that the coverage, defined as the number of websites for which they provide a meaningful label, varies enormously across providers, ranging from over 90% to below 1%. In addition, we find that all the services deviate from their documented taxonomy, hampering sound usage for research. Further, labels are highly inconsistent across providers, which show little agreement over domains, making it difficult to compare or combine these services. We also show how the dynamics of crowd-sourced efforts may be obstructed by scalability and coverage aspects and personal disagreements among human labelers. Finally, through case studies, we showcase that most services are not fit for detecting specialized content for research or content-blocking purposes. This analysis wants to bring the attention to the research community about the risk of using domain classification. Researchers should be aware of the different characteristics and deficiencies of the domain classification services to prevent the negative impact they could have on the research results and conclusions.

3. **Auditing Profiling and Ad Targeting Algorithms.** The main advantage of digital marketing over traditional advertising mechanisms is its capacity to distribute personalized ads, which rely on the information companies infer from the user's online behavior. Despite the importance of such profiles for advertisers and users, very little is known about their accuracy and the subsequent impact on ad targeting algorithms' performance. In

Chapter 5, we present an analysis of the accuracy of profiling and ad targeting algorithms from Google and Facebook, the two most relevant stakeholders in the online advertising business. We implemented a browser add-on extension that collects the user's profile on Google and Facebook and the ads received from them on general websites and Facebook social networks. In addition, the add-on allows users to rank the accuracy of each of their profile attributes. This add-on has been installed by 62 users from whom we have collected 4,311 unique profile attributes, 2,409 attribute scores through 6,400 responses in the survey, and 193,842 ads. The analysis of the collected data suggests that both Google and Facebook's profiling algorithms offer a rather low accuracy, which could seriously impact the performance of targeted ad campaigns.

1.2 DISSERTATION OUTLINE

The contributions of this dissertation, described in the above section, are structured in the document as follows.

Chapter 2 presents the principal concepts needed to understand the dissertation and its motivation. It starts with a brief description of the main tracking techniques on the web, the main aspects of new regulatory frameworks, and how they affect the operation of online services. We conclude the chapter by explaining the concept of online behavioral advertising and the components needed to operate.

The dissertation continues with Chapter 3 which includes an exhaustive behavioral analysis of pornographic websites, analyzing the presence of tracking methods and their lack of regulatory compliance, including the presence of cookie consent forms and age-verification mechanisms to avoid access to minors.

Chapter 4 presents a large-scale analysis of some of the most popular and most used domain classification services. We mainly explore their methodologies, scalability limitations, suitability for academic research, and other practical applications such as content filtering.

Chapter 5 presents a novel methodology to quantify the performance and accuracy of the user profiles generated in the online ad ecosystem and their impact on the ad targeting algorithms. These algorithms decide the ads that each user receives. The analysis focuses on Facebook and Google, two of the most relevant actors in the online advertising industry. Chapter 6 reviews the primary ethical considerations taken on this dissertation to ensure that it follows the ethical standards needed to conduct this dissertation.

Finally, the dissertation concludes with Chapter 7 briefing up the main findings and the conclusions obtained and presenting the future research lines of this dissertation in Chapter 8.

2

BACKGROUND AND RELATED WORK

In this chapter, we introduce the main concepts required to understand this dissertation and the current state-of-the-art around the topics and areas of knowledge of this dissertation. We start with web tracking and the main mechanisms implemented to track users on the web. We continue introducing the new data protection laws, focusing on the General Data Protection Regulation (GDPR) [1], which aims to increase privacy protection for all European Union citizens. We conclude by introducing Online Behavioural Advertising (OBA), which consists of a practice that consists of delivering ads to final users based on inferred users' interests.

2.1 WEB TRACKING

Web tracking consists of the techniques used to collect personal data from users when interacting with websites, including behavioral ones. The main objective of web tracking is to obtain valuable information about how users behave on the web, infer preferences, and create an online profile.

We distinguish two different types of tracking, depending on who does the tracking and where it happens. On the one hand, we define first-party tracking services when the hosting provider collects the data. Usually, the main goal behind the data collection is to understand the users' preferences, place the most relevant content in an outstanding place on the platform, or study how they interact. On the other hand, we define third-party tracking services as those embedded on websites that are not operated or controlled by the website owner. The main goal of third-party services is to provide external functionalities, including collecting insightful data from the users to create accurate online profiles to distribute tailored ads or sell them (*e.g.,* data brokers).

There are multiple tracking methods, each with its characteristics, limitations, and advantages. The following section describes the most popular and studied web tracking methods.

2.1.1 *HTTP Cookies*

HTTP Cookies, known as Cookies, are pieces of information defined as the tuple $name = value$, as the IETF establishes on its RFC6265 standard [6] and stored in the user browser. Cookies have multiple applications related to maintaining the session's state, like remembering the log-in, the shopping cart on an e-commerce site, or even preventing fraudulent activities. However, they can be used to track users and identify which websites they visit or how they interact within the website, like clicking on certain links or buttons. They are among the

most widely used tracking methods on the web and a deeply studied tracking method by the research community [7, 8, 9, 10].

Cookies Syncronization. The implementation of Same-Origin Policy (SOP) [11] on modern browsers limits the interaction between two different services with different origins on a website for security reasons. This policy establishes that the resources loaded in the browser by one origin can not access (or interact with) the resources of another origin. In addition, SOP avoids that different third-party services can share cookie IDs, so they can not merge the information they have from the same user, even when they are partners. Third-party services use a technique called Cookies Synchronization (CSync) to circumvent this security mechanism, allowing them to know the different cookie IDs assigned by different third-party services belonging to the same user. CSync allows third-party services to track users and identify which website they visit, even when they are not present on the website. For example, if just a single third-party service installs a cookie, it could inform its partners by sending them the Cookie which identifies the user. Once they receive it, they only need to check on their database which user matches the cookie ID the partner sends to them. Similarly to tracking based on Cookies, several research efforts have studied this method deeply [12, 13, 14, 15, 16, 17].

Evercookies. Evercookies follow a similar approach as cookies. However, there are fundamental differences between them. While cookies are easy to remove, and new modern browsers have started to include restrictions on installing third-party cookies, evercookies rely on multiple storage systems to identify users even when the users explicitly delete and deny the use of cookies [18]. These storage systems include the LocalStorage, Session Storage, IndexedDB, or inserting information into the HTTP Header allowing trackers to spawn or recreate removed cookies. After Soltani *et al.* [19] observed this method for the first time in 2009, more research works looked at the use of evercookies in the wild [13, 20, 19, 21, 7].

2.1.2 *Fingerprinting*

Trackers have also implemented more sophisticated and advanced techniques which allow the creation of a unique user identifier by accessing and processing several user characteristics through the execution of code in the browser. Fingerprinting techniques are distinguished from other conventional web tracking methods, like Cookies, in that they are stateless, as they collect data related to the device's characteristics. Also, they are harder to detect and prevent their execution. Furthermore, fingerprinting techniques are not homogeneous, and each exploits a specific feature that modern browsers provide. In the following, we will describe the most popular web fingerprinting techniques studied in the literature.

Canvas Fingerprinting. This technique, discovered by Mowery *et al.* [22], and initially studied in 2014 by Gunes Acar *et al.* [13], exploits the Canvas API [23] of modern browsers. This API allows drawing graphics, photos, or even animations on the browser using JavaScript. The technique consists of creating specific images in the background with a determined height, width, fonts, and background colors, among other characteristics. Then, the script responsible for creating the image transforms it to a Hexadecimal hash value that identifies the device. However, the hash value can be different despite the same script creating the same image for all the users due to the differences in how each device renders the image. The Operative System, the GPU, the graphics drivers, and the browser impact the value generated. While this technique does not identify individual users, it can conduct a unique id combined with other information. Several research works have studied the use of Canvas Fingerprinting in the wild [24, 25, 26, 27].

Other Fingerprinting Techniques. Fingerprinting techniques have evolved. Recent research works have explored other browser and device features to generate unique identifiers of the user. In this group, we can find works analyzing and exploiting vulnerabilities on the add-on extensions [28] or just the presence on the browser [20]. AudioContext is a fingerprinting technique discovered by Englehardt *et al.* [24] that relies on audio signals. This method relies on the Web Audio API [29], which provides the capacity to create audio signals. The method generates such signals and processes them to generate a user id. Similarly to Canvas Fingerprinting, each device has its characteristics that affect the audio signals, which can then be used to determine the user's device.

2.2 GDPR AND REGULATORY FRAMEWORKS

The many privacy cases of abuse inflicted by the online industry in the latest decades have motivated regulatory and legislative efforts to protect consumers' privacy and digital rights. New comprehensive data protection laws such as the European General Data Protection Regulation (GDPR) – which became effective on May 25th, 2018 [1] – aim to bring transparency to web services and empower users with control over their identity on the web and beyond. In the case of online services, this objective is achieved by forcing companies with a digital presence to obtain explicit consent from any European visitor before collecting, processing, or sharing personal data on their sites. The GDPR also gives users the right to access, correct, and delete their data collected by online services, revoke their collection consent at any time, and object to automatic data processing. Nevertheless, the GDPR will be complemented by the ePrivacy regulation once the negotiations end [30, 31]. Since its implementation, research works have studied the compliance of GDPR and the ePrivacy Directive by online services from different angles. Examples are the adoption of cookie consent banners [32], their effectiveness in providing consent [33], the analysis of the cookie banner text from a legal perspective [34] or the effects of the implemen-

tation of GDPR on the web [35, 36, 37, 17, 38, 39, 40, 37, 41], studying different changes, including the presence of third-party services or the use of tracking methods like Cookies.

2.2.1 *Sensitive personal data and pornographic websites*

The GDPR imposes additional requirements and restrictions on special categories of personal data. Article 9.1 [42] of the GDPR states that *"revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, also the processing of genetic data, biometric data to uniquely identify a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited"*. In addition, the GDPR will require website owners to obtain explicit consent from users to install and use tracking methods. The use of HTTP cookies, except when it will be strictly required to provide a service requested by the user, to fulfill a legal mandate, or to carry out certain transmissions [43]. All these requirements will be in force until the ePrivacy regulation becomes effective.

Similar regulatory efforts are taking place in other jurisdictions that used to have a traditional *laissez faire* attitude towards privacy. Notable examples are California's Consumer Privacy Act [2] (CCPA, which passed in June 2018), the Japanese Act on Protection of Personal Information [44] (effective since May 2017), and the Indian Personal Data Protection Bill of 2018 (PDP) [45]. All the regulations, as mentioned earlier, classify and consider information regarding a user's sexual life and orientation as sensitive personal data that require special treatment. Regardless of these requirements, previous research pieces of evidence provide detailed analysis of the presence of sensitive interests in users' profiles related to sexual orientation and behaviour [46].

Despite the popularity of porn websites, the online porn industry has remained largely underground. There have been isolated steps towards studying this ecosystem, mainly from a content availability standpoint in a major porn website [47] or their economic structure [48]. However, no study has deeply analyzed the regulatory compliance of pornographic websites. Only anecdotal evidence suggests the presence of tracking mechanisms [49, 50]. However, the website dataset of these studies was very limited, and the results were difficult to extrapolate to the entire ecosystem.

2.2.2 *Access Control in Pornographic Sites*

For two decades, many laws failed to effectively prevent children from viewing pornography and other harmful materials, including pornographic content, on the Internet [51]. Several efforts have taken place in the United Kingdom. The 2017 Digital Economy Act [52], which was delayed twice before becoming abandoned in October 2019 [53], aimed at enforcing the deployment of age verification mechanisms to block minors from accessing pornographic material. Nevertheless, the UK government has brought back the idea of minor control access to adult material by implementing the Online Safety Bill [54, 55], which

will become effective sometime in 2022. Similar efforts are taking place in other European state members like France, where the authorities have implemented new regulations to prevent minors from accessing porn websites [56, 57].

To comply with the new age-verification laws, the industry designed and developed tools such as AgeID, a technology proposed by MindGeek [58] that is expected to become an industry standard [59]. A complementary effort to the aforementioned methods is the proposal made by the Association of Sites Advocating Child Protection (ASACP) [60]. This not-for-profit organization has created a Restricted-for-Adults (RTA) meta tag to assist parents in preventing their children from accessing pornographic material. The fact that there are companies from the online porn industry among the members [61] of this association is considered a good example of collaboration between the porn industry and external organizations to increase safety and regulatory compliance.

Other regions in the world have followed more drastic and polemical strategies. For example, the Russian government requires Pornhub users to log in with a social network profile linked to their passport number [62, 63]. This measure has raised several ethical and privacy concerns. In addition, world countries like most Middle East countries, India, Iran, and China actively ban, prosecute, and prohibit access to pornographic content altogether [64, 65]. The 2013 Anti-Pornography Act in Uganda prosecutes the broadcasting and trading of pornography [66], while the Anti-Homosexuality Bill Act in 2014 prosecutes LGBTI communities [67].

Previous research works have analyzed the use of censorship techniques to prevent citizens from a country from getting access to certain types of websites, including pornographic websites [68, 69, 70]. However, little is known about the implementation, the limitations, and the privacy problems that access control techniques in porn websites face users.

2.3 ONLINE BEHAVIOURAL ADVERTISING

We have mentioned in the previous chapter that online advertising plays a pivotal role in the business model for many online services, as it has become one of their main funding sources. Online advertising has grown and evolved in the last years, gaining market share over more traditional advertising media platforms, like TV or radio, due to its capacity to reach users with specific characteristics for a low price.

Online advertising has developed a complex profiling ecosystem that exploits the data users generate when interacting with the websites using tracking techniques (see Section 2.1) to infer the users' preferences. All this rich information serves as input data to advertising companies' profiling algorithms. These sophisticated algorithms can map the actions of users into preferences or interests on certain topics through tagging systems. Additionally, the algorithms use other non-online activity information, like demographic information, that users directly provide (*e.g.,* users can provide personal data when they create an account on Google). Finally, the advertising services create the profiles combining both types of information. The online advertising ecosystem is interested in cre-

ating accurate profiles of the users and adjusting the ads based on such profiles because it increases the probability that users will click and maintain the ad and the product announced in the future.

When a user visits a website, mobile app, or any other online service with the capacity to show ads, the ad space (typically embedded in an iFrame [71]) sends a request to the entity handling it (*e.g.*, Google) called the ad network. This entity is responsible for taking care of the ad space, compiling all the possible information about the ad space: 1) information of the space itself (website, size, allowed type of ad, position on the page, among other information.); 2) information about the browser, operating system, and type of device (mobile vs. desktop); 3) information about the user visiting the website, *i.e.*, the user's profile. At this point, the ad network looks for an advertising campaign whose audience (the users with specific preferences they want to reach) matches the user's profile from among the advertisers configuring their campaigns on its platform. Typically, there is more than one campaign matching the offered profile. Then, it runs an auction process to choose the ad campaign whose ad will be delivered.

The process required to distribute personalized ads occurs in an opaque way, from the profile creation to the decision to distribute an ad to a specific user. Nevertheless, several studies have made groundbreaking contributions to shed light on the ecosystem created. Mainly, previous works have studied the information included on the profiles that big tech firms have inferrer, including the presence of sensitive attributes that later use to distribute ads [46, 72, 73, 74]. In this context, the research community has quantified the prevalence of online targeting advertising in the online advertising ecosystem [75, 76]

Part II

A COMPREHENSIVE PRIVACY ANALYSIS OF THE WEB PORN ECOSYSTEM

3

A COMPREHENSIVE PRIVACY ANALYSIS OF THE WEB PORN ECOSYSTEM

Pornographic (porn) websites are among the most visited and lucrative online services since the early days of the World Wide Web [77]. Pornhub, the most visited porn website according to Alexa's domain rank [78], had 33.5 Billion visits and was returned in 30.3 Billion web searches in 2018 [79]. MindGeek, Pornhubs' parent company, has reported over half a billion dollars of revenue in the 2015 fiscal year [80].

Modern privacy regulations like the EU General Data Protection Regulation (GDPR) [42] and California's Consumer Privacy Act (CCPA) [2] consider sexual information of an individual as highly sensitive data. All these privacy regulations require organizations with an online presence to request informed consent from users prior to any data collection [44, 45, 2, 42]. However, as in the case of regular websites, pornographic ones also integrate third-party components – *e.g.,* advertising and analytics libraries – with the capacity to track users' interaction with such services and, therefore, potentially infer a visitor's sexual orientation and preferences.

The collection of this information, in addition to the absence of secure network protocols like HTTPS, could put at risk visitors of those websites, specially those connecting from countries where certain sexual orientations are prosecuted [67, 81, 63, 82, 83].

Despite the many research efforts that have taken place in the last decade to identify and quantify the presence and use of tracking technologies in the web, no study has deep dived yet into the privacy risks of sensitive websites, like pornographic ones. It is unclear, as a result, whether pornographic websites can pose a privacy risk to their visitors, and if they comply with the provisions set both by privacy regulations and by new rules at the time this works was conducted, to control minor's access to adult content like the UK's Digital Economy Act [52][1]. In fact, anecdotal evidence suggests that there are significant differences between the third-party organizations operating in the porn and the regular web tracking industry [49] as large online ad networks such as Google Ads set strict constraints for porn-related publishers, prohibiting the advertising of adult-oriented products and services [82]. These restricting terms of services – possibly driven by fear of damaging their brand reputation – opened new market opportunities for other actors who have specialized in providing advertising and tracking technologies to adult sites. This context has created, as a result, a parallel ecosystem of third-party service providers in the porn ecosystem who has not been scrutinized by regulators, policy makers, and the research community.

1 The implementation of the UK's Digital Economy Act was delayed twice before becoming abandoned in October 2019, just after the works was published and presented at ACM IMC 2019

We develop and used a methodology to perform the first holistic analysis of pornographic websites from a privacy, transparency, and regulatory compliance perspective. Our main contributions are:

1. We design a semi-supervised method to compile a representative sample of pornographic websites using publicly available resources. After manually inspecting and removing false-positives, we identify 6,843 different pornographic websites.

2. We develop and use a methodology to study the presence of third-party services in the porn ecosystem. We compare the presence of third-party services present in pornographic websites with those embedded in the most popular web sites according to Alexa's rank. We find 3,673 third-party services embedded in porn websites, including companies specialized in the porn industry (*e.g.,* ExoClick), well-known advertising companies (*e.g.,* DoubleClick), analytics services (*e.g.,* Google Analytics), and domains associated to data brokers (*e.g.,* Acxiom). 84% of the third-party services embedded on pornographic websites do not appear in the most popular non-pornographic websites.

3. We study the behavior of pornographic websites and the third-party tracking services embedded in them. We find the presence of third-party HTTP Cookies in 72% of the analyzed pornographic websites, while 5% of them also use advance fingerprinting techniques like Canvas Fingerprinting to identify visitors uniquely. Interestingly, 91% of the scripts we found using canvas fingerprinting are not indexed by EasyList and EasyPrivacy [84].

4. We quantify behavioral differences on porn websites depending on the user's location and jurisdictional area. We conclude that the number of third-party services is quite stable across countries, yet there are regional third-party services that only operate in specific regions: *e.g.,*27 ATS only appear in Russia.

5. We develop and validate a method to automatically analyze the transparency and regulatory compliance of pornographic websites. Specifically, we study the presence of cookie consent banners, privacy policies, and age-verification mechanisms. Our analysis reveals a significant absence of privacy policies and consent forms across pornographic websites in spite of their sensitivity. This pattern holds even in regions with strict regulatory frameworks like the European Union: only 16% of the websites have privacy policies when accessed from a machine located at a EU state member. Finally, only 4% of the analyzed porn websites implement cookie consent forms.

Our study reveals a concerning lack of transparency in pornographic websites, despite the large presence of third-party trackers embedded in them and an increasing regulatory pressure. Therefore, we believe that our study will contribute to stress the importance of studying subsets of the world wide web that offer sensitive services and content in depth. This type of effort is not only

needed to effectively inform the privacy debate, but also to promote user awareness.

3.1 DATA COLLECTION AND METHOD

The first challenge in our study is compiling a representative list of pornographic websites. For that, we implement a semi-supervised approach that combines three different data sources and steps with varying levels of accuracy:

1. We combine all the pornographic websites indexed by three websites specialized in aggregating, recommending, and classifying pornographic content [85, 86, 87]. This process provides us with 342 porn websites.
2. We extract 22 websites classified as *Adult* sites by the Alexa's website categorization service [88].
3. We look for websites indexed by Alexa's rank [78] (throughout 2018) that are potentially offering pornographic content by searching for keywords related to pornographic and adult content in their URLs (*e.g.,* "porn", "tube", "sex", "gay", "lesbian", "mature" and "xxx"). We find 7,735 websites matching these substrings.

The combination of these three methods allows us to identify 8,099 potential pornographic websites. However, the third keyword-based method introduces false positives if not done with care, since the chosen bag of words is not exclusively related to pornographic material (*e.g.,* PornTube offers pornographic content while YouTube does not). To identify and remove false positives, we implement a purpose-built crawler to download their content (DOM and screenshots) which are then manually inspected. In total, we find 1,256 false positives, many of which are because of unresponsive websites at the time of the crawl (we investigate below the stability of these domains). After this sanitation process, we obtain a corpus of 6,843 pornographic websites of various kinds, including websites hosting user-uploaded videos and live streaming content, or websites acting as proxies to pornographic material (*e.g.,* `pornsource.com`), among others. Finally, we use a reference dataset containing 9,688 popular non-pornographic websites[2] to study the commonalities and differences between sensitive pornographic websites and regular ones.

Popularity of Pornographic Websites: We use a longitudinal dataset containing the Alexa top-1M sites throughout 2018 as a proxy to measure the stability, popularity, and representativeness of our corpus of pornographic websites. Figure 3.1 shows the best and median rank value for each one of the identified porn websites, as well as the percentage of days each website was in the Alexa top-1M over the whole year.[3] We find that 1,103 websites (16%) were always present in the Alexa top-1M, and just 16 of them were always within the top-1K websites during the one-year period (*e.g.,* `pornhub.com`, `xvideos.com` or `livejasmin.com`).

2 Websites extracted from Alexa's top-10K, in the 10th of January 2019.

3 We consider their popularity for a whole year in order to account for any eventual bias caused by the Alexa ranking [89].

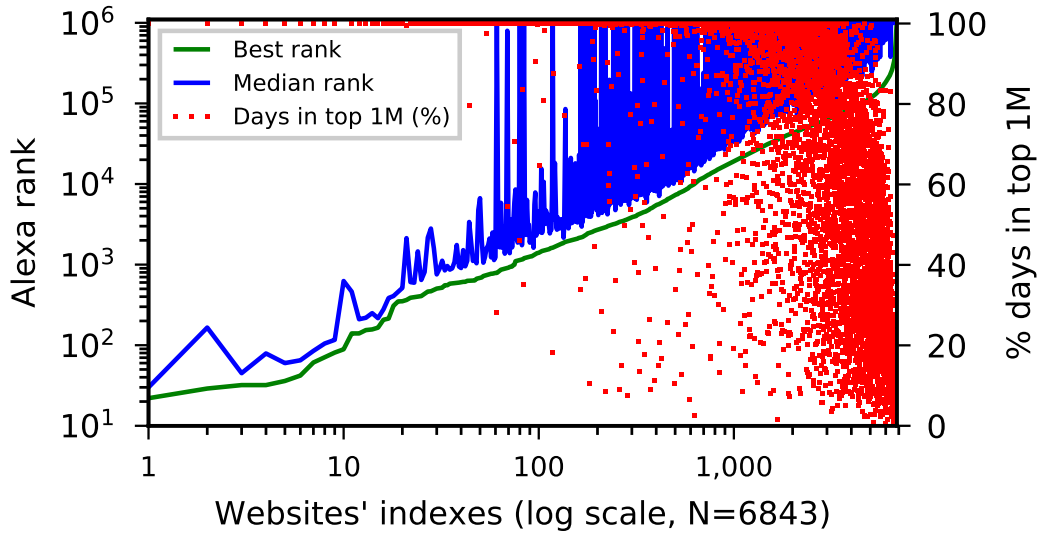


Figure 3.1: Best (green) and median (blue) Alexa rank for each pornographic website, and the percentage of days that each one of them were indexed in the top-1M throughout 2018. The pornographic websites are ordered in the x-axis by their best Alexa rank.

3.1.1 *Web Crawlers*

Our analysis and data collection workflow uses two complementary crawlers to study the behavior of pornographic websites as shown in Figure 3.2. First, we use a OpenWPM-based crawler to collect evidence of the behavior of each website and used tracking technologies, as well as the presence of third-party libraries and privacy consent forms. In both cases, we only crawl the landing page of websites so our study presents a lower-bound estimation of the privacy risks of pornographic websites as we do not interact with them beyond their landing page. Second, we use a Selenium-based crawler to automatically interact with each pornographic website to pass through the age verification mechanism (when available) and collect the privacy policy. We provide further details about each crawler and their purpose below.

OpenWPM: Rather than implementing yet another crawler, we use OpenWPM [24] because of its simplicity, stability, and the versatility of the features that it offers. OpenWPM is based on Firefox version 52 and allows (1) collecting all the HTTP and HTTPS requests and responses generated while crawling a website; and (2) detecting different tracking technologies, including advanced ones like canvas fingerprinting [24]. Nevertheless, we extend OpenWPM capabilities to analyze other other aspects of pornographic websites. First, we develop methods to extract the chain of requests caused by Real-Time Bidding (RTB) processes (*i.e.*, the inclusion chain [16]) to identify third-party services dynamically embedded in the target websites [16]. Specifically, we analyze the HTTP `Referrer` headers and remove those third parties not directly called by the publisher. Finally, we also enable mechanisms in OpenWPM to automatically record both HTTP cookies and cookie consent forms, so that we can estimate the transparency and regulatory compliance for each pornographic website (Section 3.5). We use the

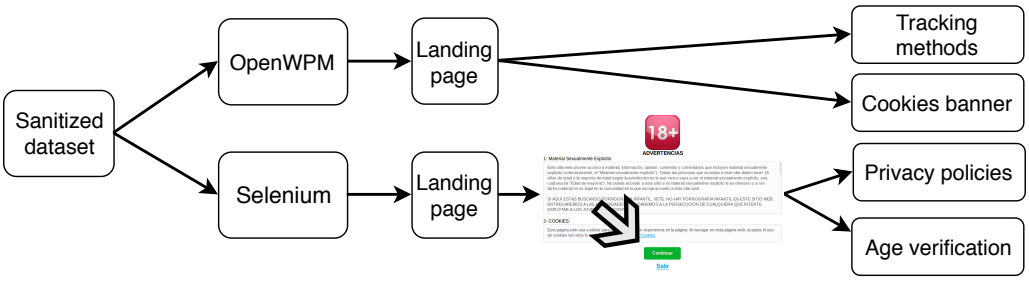


Figure 3.2: Workflow of our data collection.

same browser session – *i.e.,* we do not close and open the browser between visits – for the duration of the crawling process, in order to be able to capture cookie synchronizations (Section 3.3.1.2).[4] It is also important to note that we only crawl each page once, giving us a lower bound on tracking activities [90].

Selenium: We implement a second purpose-built Selenium-based Chrome crawler to (1) detect and bypass age-verification mechanisms in pornographic sites; and (2) fetch their privacy policies when available. We separate this data collection process from OpenWPM crawls to avoid any instrumentation bias introduced by the need of interacting with each website in order to identify their privacy policies. To detect and quantify the support for age verification mechanisms, our crawler parses the landing page of a website and searches for floating elements and the words "Yes", "Enter", "Agree", "Continue" and "Accept" in 8 languages. We select English, Spanish, French, Portuguese, Russian, Italian, German, and Romanian for being the most common default languages in our list of pornographic websites. We choose these keywords after a manual inspection of part of the websites in our corpus. To eliminate false-positives introduced by using keyword-based matching, our crawler inspects the HTML DOM and the text of the parent and grandparent elements of those containing any of these keywords to identify and verify the presence of age verification mechanisms or warning messages about the content of the webpage. If a relevant message is found, then the crawler clicks on the element to access the landing page. Finally, we fetch privacy policies by searching for URL links containing the keywords "Privacy" and "Policy" in any of the 8 languages. We manually validate the accuracy of our method in Section 3.5.2.

Geographical diversity: One of the goals of this work is to study whether pornographic websites behave differently depending on the user location and jurisdiction. To answer that, we run our crawls from a vantage point located in Spain, and use two commercial VPN providers – NordVPN [91] and PrivateVPN [92][5] – to gain access to vantage points in other EU state members, Singapore, India, Russia, USA, and the UK [6]. When crawling from Russia and India, we could

4 We established a timeout of 120s for loading a website in order to prevent our crawlers from becoming stagnant.

5 We select those VPN providers because 1) they do not appear to manipulate traffic according to our experiments, and 2) they forward traffic through VPN servers rather than through real users in a P2P fashion [93].

6 We perform these measurements in the UK to study websites' compliance with the Digital Economy Act [52]

not access 21 and 168 pornographic websites, respectively. Unfortunately, we can not assert whether this is due to country-level censorship or server-side blocking [94].

3.2 THE PORN WEB ECOSYSTEM

As of today, the research community lacks of generalizable and robust methods to classify domains by the type of service that they offer, and to identify their parent company [95]. However, gaining this knowledge is critical not only to identify websites offering sensitive content and to be able to identify the organization providing tracking services, but also to assess the accountability of these organizations. In this section we explore (1) the main organizations providing pornographic content and their business models (Section 3.2.1); and (2) the analysis of third-party tracking technologies embedded in pornographic sites (Section 3.2.2).

3.2.1 *Discovering Website Owners*

Discovering the parent company or organization supporting a given website is a hard problem that requires applying complementary methods. We start this analysis by crawling and measuring differences across pairs of pornographic websites at the landing page and privacy policy (when available) of each pornographic website to search for organization-level information. For the majority of pornographic websites, this information is either vague or incomplete: *e.g.*, some websites only report a postal address rather than a company name accompanied by legal information. Second, we apply the term frequency-inverse document frequency statistical method (TF-IDF) [96] to measure the similarity between privacy policies and the HTML `<head>` element of each pair of pornographic websites to automatically find clusters that might belong to the same organization. We manually analyze each pair and cluster to remove potential false-positives. This method allows us to identify over 80 porn websites that belong to six different companies, including AFS Media LTD., Techpump, Gamma Entertainment, and PaperStreet Media. To increase the coverage and improve the accuracy of our attribution process, we leverage DNS, WHOIS, and X.509 certificate information and insights obtained from white papers, scientific articles, and public reports about the pornographic industry [97, 98].

The combination of these methods only allows us to accurately find 24 companies owning 286 pornographic websites. We could not find reliable organization-level information for 96% of the pornographic websites in our dataset. This lack of corporate or organizational transparency is particularly concerning for websites – data controllers in the context of GDPR – engaging in user tracking or embedding third-party services as their visitors will not be able to effectively exercise their privacy rights to any corporation (*e.g.*, demanding access, corrections, or deletion of their data as indicated in the GDPR). We further discuss in Sections 3.3 and 3.5 the presence of trackers in pornographic websites, and their long way towards regulatory compliance, respectively.

Company	# sites	Most popular site (rank)
Gamma Entertainment	65	`evilangel.com` (5,301)
MindGeek	54	`pornhub.com` (22)
PaperStreet Media	38	`teamskeet.com` (10,171)
Techpump	25	`porn300.com` (2,366)
PMG Entertainment	15	`private.com` (7,758)
SexMex	12	`sexmex.xxx` (122,227)
Docler Holding	10	`livejasmin.com` (36)
Mature.nl	9	`mature.nl` (6,577)
Liberty Media	7	`corbinfisher.com` (26,436)
WGCZ	5	`xvideos.com` (32)
AFS Media LTD	5	`theclassicporn.com` (13,939)
AEBN	5	`pornotube.com` (31,148)
Zero Tolerance	5	`ztod.com` (40,676)
Eurocreme	5	`eurocreme.com` (110,012)
JM Productions	5	`jerkoffzone.com` (147,753)

Table 3.1: Largest clusters of pornographic sites, grouped by their parent company. For each company, we report the number of individual websites owned and the one with the highest Alexa rank throughout 2018. A larger cluster size does not necessarily translate into popularity.

Main pornographic website operators: Table 3.1 shows the 10 largest clusters of organizations ordered by the number of individual pornographic websites that they own. These companies own and operate 3% of the total websites in our corpus. The reasons behind these clusters or pornographic websites are manifold. Typically, these clusters are created through acquisitions and mergers between companies, similar to the industry trends present in the online advertising and tracking industry [95, 99]. Furthermore, pornographic websites are typically federated. This gives them the ability to reach out larger audiences and increase advertising revenues through affiliated services, while also re-publishing and sharing pornographic material across sites.

Monetization Models: The majority of pornographic websites combine different monetization mechanisms, such as online advertising-based models (see Section 3.2.2), subscription (premium) services, and, in some cases, even through cryptomining services (see Section 3.3.3). We perform a semi-automatic classification of these websites to infer their business models. First, we parse the landing page of the websites in our dataset and look for keywords that may indicate the option to create an account (*e.g.,* "Log In", "Sign Up") or "Premium" services. We use this signal as a proxy to identify which websites may offer subscription-based services after authentication. Then, we manually label the subscription model as "free" (*i.e.,* the content is freely available after registration), or "paid" (*i.e.,* the content is protected by a payment wall) by inspecting

Domain category	Pornographic websites (P)	Regular websites (R)	$\lvert P \cap R \rvert$
Corpus size	6,346	8,511	—
First-party	727	3,852	—
Third-party	5,457	21,128	889
Third-party ATS	663	196	86

Table 3.2: Number of first party and third-party domains found on our dataset of pornographic and regular websites. ATS makes reference to third-party Advertisement and Tracking Services.

the website. We also verify that the keywords for creating an account and for detecting premium services remain stable independently of the language of the webpage. Thanks to this method, we can conclude that 14% of the porn websites in our corpus offer subscription options; and only 23% of the websites require a payment. While the study of the privacy risks of subscription-based services is outside the scope of this paper, it may be possible that once a user creates an account, all of their actions might be also linked to their profile and banking information.

3.2.2 *Third-Party Services in Porn Websites*

A large number of pornographic websites rely on online advertisements to monetize their user base and content and on analytics services for tracking their audiences. However, many ad networks set strict limitations on the usage of their services in pornographic websites, possibly as a measure to protect their brand reputation [82]. This state of affairs has given birth to lesser known ecosystem of advertisement and tracking services (ATS) specialized in adult content which have escaped research and regulatory scrutiny. We conjecture that our current limited understanding of trackers in sensitive websites has been caused by the low penetration of some of these trackers across the whole web landscape, hence falling in the long-tail. In fact, many pornographic websites are rarely indexed in domain ranks so they might not be present in studies that crawl a one-day sample of popular domain ranks [89].

In this subsection we study the third-party services and organizations operating in the online porn industry, and compare them with those present in regular websites. With our OpenWPM-based crawler, we find 5,457 different third-party domains embedded in the set of 6,346 pornographic websites that we could successfully crawl (out of our 6,843 sanitized dataset of pornographic websites). An eyeball analysis of these domains reveals that the majority of them belong to third-party analytics and advertising services, but also to CDN providers and social networks. To obtain a more accurate picture of the third-party tracking ecosystem in pornographic websites, we use the following complementary heuristics to (1) label and classify the domains embedded in pornographic web-

sites as first-, third-party, or third-party advertising and tracking (ATS) services; and (2) attribute hostnames to organizations:

1. **Third-party service extraction:** We collect all the URLs from all the HTTP(S) requests triggered by our OpenWPM-based crawler to identify the presence of third parties. For comparison, we run our crawl both for our pornographic and regular website datasets. For each URL and HTTP(S) request, we compare its fully qualified domain name (FQDN) and its X.509 certificate information (when available) along with the FQDN and certificate information of the host website, to determine whether a service is a first or third party. If we cannot establish a relationship between a host website and an embedded service based on the previous method, we compute the similarity between the two FQDNs using the Levenshtein distance [100]: if the similarity is higher than 0.7, we then consider the FQDNs to belong to the same entity. We manually verified the results and found this method to be accurate. This method also allows us to group together domains such as `doublepimp.com` and `doublepimpssl.com`, but also to make the distinction between *e.g.,* `doublepimp.com` and `doubleclick.net`. We can successfully label as third party domains 91% of the 6,017 FQDNs contacted when crawling all the porn websites by using this technique.

2. **ATS classification:** We rely on EasyList and EasyPrivacy blocklists [84] – downloaded on Jan. 29th, 2019 – to identify domains belonging to well-known ATSes. These blocklists are designed and used by the AdBlock [101] and AdBlockPlus [102] browser extensions. Since they are based on rules that consider the whole URL request (*e.g.,* `bbc.co.uk` is not blocklisted, but `bbc.co.uk/analytics` is), we match the full URL provided by OpenWPM with these blocklists to identify actual instances of tracking. We relax the matching method to the base FQDN domain to identify the presence of 12% third-party ATS organizations [103].

3. **Finding the parent company for third-party services:** To better understand the trackers and organizations involved in the ecosystem, it is also critical to associate third-party domains to their parent company. We initially considered using Disconnect's domain-to-company mapping [104] but we soon realized that it is incomplete. We designed a method to complement Disconnect's list with organization-level information found in the X.509 certificate of each third-party domain[7], hence improving significantly its accuracy and coverage. For instance, we could assign to Oracle several third-party trackers like `addthis.com` (AddThis) [105] and `bluekai.com` (BlueKai) [106] services [8]. After this process, we found the parent company for 4,477 (74%) FQDNs, accounting for 1,014 companies, while using Disconnect's list yields only 142 of them.

7 In some cases, the `Subject` field only contains the domain name of the website instead of the company name. We choose not to take the certificate information of these websites into account.

8 Oracle operates a data marketplace, the largest third-party data marketplace for "open and transparent audience data trading" according to their own sources [107].

3.2.2.1 *Third-Party in regular versus porn websites*

Table 3.2 compares the number of third-party domains present in our set of pornographic websites with those present in our reference set of regular websites. This comparative analysis uncovers significant differences. In aggregated terms, we found 21,128 third-party domains (FQDNs) in our set of regular websites but only 5,457 in the pornographic ones. However, when looking specifically at ATS services, we see that they are more widespread and diverse in pornographic websites as 12% and 1% of all the third-party domains found in pornographic and regular websites are associated with ATSes, respectively. The intersection between the set of ATSes operating in the regular and pornographic websites is also low: only 86 third-party advertising and tracking services are present in both types of websites. This analysis reveals that a majority of advertising and tracking services operating in the online pornography ecosystem are unlikely to be present in regular websites. For instance, `exosrv.com` and `exoclick.com`, both belonging to Exoclick, are found in 2,709 pornographic websites (43% of the corpus) but only in 6 regular websites. These figures only represent a lower-bound estimation of the presence of advertising and tracking services in pornographic websites due to the well-known limitations of existing domain classifiers and blocklists [95, 108]. In Section 3.3, we will inspect the behavior of each third-party service to identify more trackers.

3.2.2.2 *A closer look at the long-tail*

The set of third-party services present in pornographic websites varies with the popularity of the hosting site. More concretely, the more unpopular the pornographic website is, the more obfuscated and opaque are the third-party domains it embeds.

Table 3.3 shows the presence of third-party services in porn websites when grouped in different popularity intervals (according to their highest Alexa rank throughout 2018). Only 3% of third-party domains, regardless of their purpose, are present in the four different tiers of popularity. Amongst those we find cloud providers such as `cloudflare.com` and large advertising companies (*e.g.,* DoubleClick by Alphabet), but also ATS companies specialized in adult websites such as `doublepimp.com` or `exoclick.com`. We would like to stress that Alphabet Inc. has specific policies about the type of content that can be distributed through their ad network as well as on the hosting site [82].[9]

In order to get a better understanding of the implications of low popularity – and possibly reputation – in terms of third-party services embedded in porn website, we take a deeper look at 2,069 unpopular pornographic sites that never got indexed by the Alexa Top 100K rank throughout 2018. This detailed analysis confirms that it is more likely to find advertisement and analytic services that are not commonly used by the prominent websites in unpopular pornographic websites. In fact, we find that 18% of the third-party services embedded on all porn websites appear only in the less popular ones

9 Performing an analysis on whether the host sites are in compliance with Google Ads Policies is outside the scope of this paper.

Popularity Interval	Number of porn websites	Third-party domains (Unique to the interval)
0 — 1k	73	407 (119)
1k — 10k	536	1,327 (531)
10k — 100k	3,668	3,702 (2,115)
100k+	2,069	2,363 (1,007)

Table 3.3: Third-party presence by popularity interval (per Alexa's 2018 highest rank). For each interval we show the total number of third-party domains ("Total") and the third-party domains found only in this interval ("Unique")

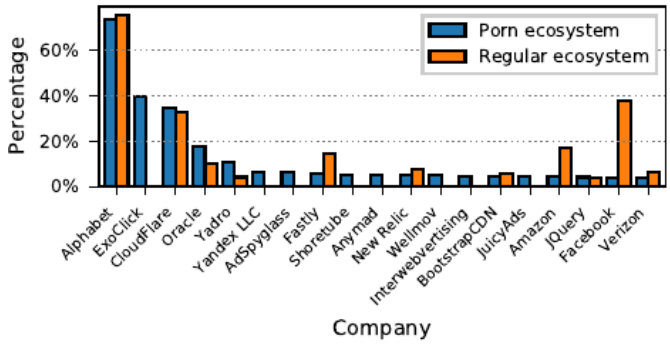


Figure 3.3: Most relevant third-party organizations in the porn ecosystem. We show their prevalence in the regular ecosystem for comparison.

according to Alexa. This is the case of analytic services like `adultforce.com` and `zingyads.com` [109, 110], for which we could not find a privacy policy on their homepages. We also found four Russian tracking services (`betweendigital.ru`, `datamind.ru`,`adlabs.ru` and `adx.com.ru`) on `pornovhd.info`, a Russian porn website. Finally, we remark the presence of a potentially malicious domains (according to Dr. Web) such as the traffic trade webpage `itraffictrade.com` [111].

3.2.2.3 *An organization-level analysis*

We now present an organizational level analysis of the third-party domains operating in pornographic websites, regardless of their role. Figure 3.3 shows the 19 companies offering third-party services to most of the studied pornographic websites. As we can see, Alphabet is – as in the regular web – the most prevalent organization (74% of the total pornographic websites). Exoclick and Cloudflare services[10] are second and third with 40% and 35% of prevalence, respectively. When comparing with the third-party companies present in the regular web, we find that several ones solely operate in the adult industry. While some of them are well-known actors like Exoclick [112], others are lesser known companies like JuicyAds (4%) [113] and EroAdvertising (4%) [114].

10 In this specific case, we cannot confidently confirm that Cloudflare is operating these domains. It might be possible that other companies, advertising services or tracking services might be using Cloudflare's infrastructure.

In general terms, the presence of Alphabet services (*e.g.,* Doubleclick, Google Analytics) is very similar in both regular and pornographic websites. Yet, the prevalence of each individual service varies greatly: `google-analytics.com` is present in 39% of porn websites, while `doubleclick.net` – an ad-network – appears in 12% of them (for reference, 60% of the analyzed non-porn websites connect to Doubleclick domains). The higher presence of Oracle in porn websites is caused by its `addthis.com` service, which provides web developers features like social network integration and content sharing (*e.g.,* pictures or videos). Another interesting case is the domain `alexa.com` which is related to the Amazon-owned browser extension that populates such list. Another interesting case is the presence of the domain `rlcdn.com` in four pornographic websites, one of them offering *bestiality porn*, a practice considered illegal in many countries of the world. This domain belongs to RalpLeaf which is a subsidiary of TowerData/Acxiom [115], one of the largest data brokers in the world [116, 117]. Finally, while Facebook is highly popular in the web ecosystem, its presence in pornographic websites is really low.

3.3 PRIVACY RISKS

The sensitive nature of pornographic websites, and the quite unique ecosystem of third-party ATSes operating in them highlight the importance of studying in depth the behavior of these websites and their use of tracking technologies. In this section, we perform a multi-dimensional analysis of the various privacy risks to which visitors of pornographic websites might be exposed to (Section 3.3.1). We also provide an analysis of the use of insecure protocols (*e.g.,* HTTP) who may allow in-path observers like censors to monitor users' browsing habits (Section 3.3.2), and report on the presence of known malware in these sites (Section 3.3.3).

3.3.1 *User Tracking Techniques*

We leverage our customized version of OpenWPM to measure the use of various tracking techniques in pornographic websites, specifically HTTP cookies, cookie syncing, and advanced fingerprinting techniques.

3.3.1.1 *HTTP Cookies*

Online companies often use HTTP cookies as a means for tracking users across the web. They do so by generating and storing unique identifiers in end-users' browsers. Using OpenWPM, we can successfully identify 89,009 HTTP cookies installed by 92% of our dataset of porn websites. This includes both first- and third-party cookies. However, not all cookies might be used for the purpose of tracking users (*e.g.,* session cookies). Therefore, we focus our analysis in those HTTP cookies that may potentially contain user identifiers. For that, we discard session cookies and those with a length below 6 characters which are unlikely to contain unique identifiers [118]. After applying this filter, 51,648 HTTP cookies

Third-party domain	% porn websites	# Cookies	ATS	In web ecosystem	% Cookies with user IP
exosrv.com	21%	2095	✓	✓	85%
addthis.com	17%	1289	✓	✓	0%
exoclick.com	14%	434	✓	✓	29%
yandex.ru	4%	312	✓	✓	0%
juicyads.com	4%	475	✓	✓	0%

Table 3.4: The 5 most common third-party domains delivering cookies that potentially contain unique IDs.

that can potentially be used for tracking users remain.3% of them are larger than 1,000 characters, even reaching 3,600 characters in the case of cookies installed by third-party ATS services like juicyads.com, tsyndicate.com, exoclick.com, exosrv.com, and other porn websites.

We now focus our study on the 30,247 HTTP cookies installed by 3,343 third-party domains in 72% of our corpus of pornographic sites. The 100 most popular cookies (by their unique $name = value$ combination) appear in over 30% of the total porn websites. Moreover, as shown in Table 3.4, the main third-party services responsible for installing HTTP cookies in users' browsers are ExoClick, Oracle (*via* AddThis), Yandex, and JuicyAds. While ExoClick and JuicyAds are specific to the online porn ecosystem, AddThis and Yandex are commonly found in regular web services, allowing these firms to potentially track users across the whole web.

Encoded Information in HTTP Cookies. We decode the cookie values using two types of encoding: base64, and URL. We detect 2,183 cookies that store the IP address of our physical machine along with potential IDs. 97% of these cookies belong to different Exoclick domains, which are present in 440 different porn websites as shown in Table 3.4. In particular, 85% of exorsrv.com cookies and 29% of exoclick.com cookies follow this pattern. Furthermore, we identify 28 cookies in 15 websites that store approximate geolocation data, potentially obtained through geo-IP databases [119]. 27 of these 28 cookies are delivered by two third-party domains, fling.com and playwithme.com. While the former only stores the coordinates, the latter also includes detailed information about the network provider. While the accuracy of geo-IP databases is not very precise in general, it could reveal the precise location of a user in certain scenarios [120].

3.3.1.2 *Cookie Synchronization*

For security purposes, modern web browsers limit the access to cookies to the service that has installed them [11]. To circumvent this security mechanism and ease cross-site tracking, third-party services use a technique called cookie synchronization (cookie syncing, in short) that allows them to share their cookie data with other services by embedding the cookie in the URL [121, 12]. We

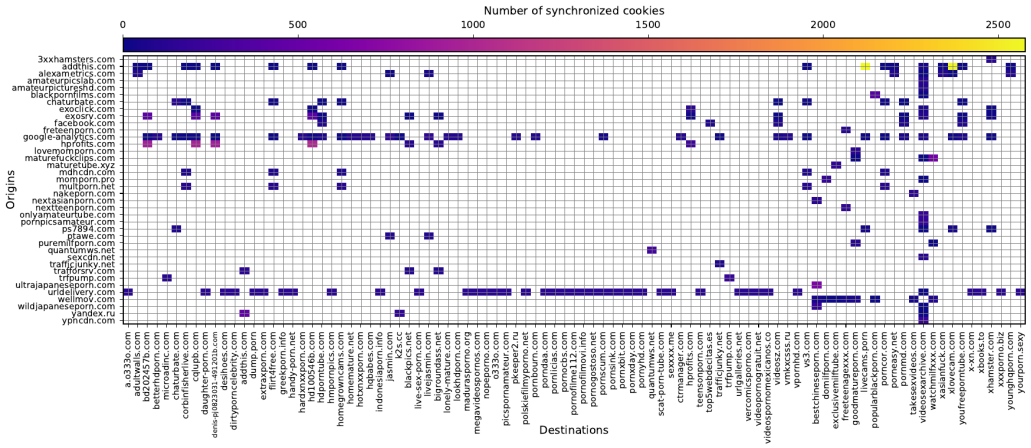


Figure 3.4: Cookie syncing between organizations. Pairs of domains that exchanged at least 75 cookies are shown.

study the use of cookie syncing in pornographic websites by checking if any of the observed HTTP cookies are later embedded in subsequent HTTP requests. To avoid introducing false positives, we do not split the cookie value by delimiters like "-" or "=". Hence, our findings offer a lower bound estimation of the prevalence of this technique.

The number of pornographic websites for which we have observed this practice is 2,867. This covers 58% of the top-100 most popular porn websites according to Alexa. However, the matching of the pairs of organizations (at the domain level) involved in this practice yields 4,675 different pairs as shown in Figure 3.4 (for clarity reasons, we only show the pairs of domains that exchange at least 75 cookies). Specifically, we can find 1,120 origin and 727 destination services. Cookie syncing can also occur between domains belonging to the same organization. For instance, the third-party domains `hd100546b.com` and `bd202457b.com` synchronize HTTP cookies with `hprofits.com`. The X.509 certificates for these three domains suggest that all of them belong to `hprofits.com`, an ad exchange platform according to their website.

3.3.1.3 *User Fingerprinting*

Fingerprinting techniques allow trackers and services to create a unique user identifier by accessing and processing several characteristics of the user's device using JavaScript APIs. As opposed to cookie-based tracking, this sophisticated method can be used to persistently track users and their activities across websites without having to rely on cookies.

First, we analyze pornographic websites and third-party services using either canvas or canvas font fingerprinting techniques [24]. HTML Canvas Fingerprinting is a tracking technique that exploits system differences between devices in how they render images. These scripts use the `CanvasRenderingContext2D` and the `HTML- CanvasElement` JavaScript APIs to generate images using specific height, width, fonts, and background colors, among other characteristics. Font fingerprinting, instead, is a variation of canvas fingerprinting in which a tracker

Domain	Presence in porn sites	ATS	Regular web	Canvas fingerprinting	WebRTC
adsco.re	152	-	✓	0	1
ero-advertising.com	33	✓	✓	32	0
cloudfront.net	31	✓	✓	8	0
cloudflare.com	28	✓	✓	2	0
adnium.com	26	✓	-	41	0
highwebmedia.com	22	✓	✓	1	0
xcvgdf.party	18	-	-	18	0
provers.pro	15	✓	-	1	0
montwam.top	13	✓	-	25	0
dditscdn.com	10	✓	✓	1	0

Table 3.5: Third-party domains using different tracking-techniques. The ATS and Regular web columns indicate whether these services are indexed in EasyList/EasyPrivacy or if they are present in the regular web, respectively.

can leverage the fonts that each browser has installed to generate a unique ID of the device. This is achieved with the measureText method of the HTML Canvas API which allows to draw text using different fonts. Depending on the size of the written text, the tracking service can infer if a particular font is installed.

Yet, not all the services that invoke these JavaScript APIs do so for the purpose of tracking users. To eliminate false positives, we follow the methodology proposed by Englehardt *et al.* [24]. In the case of canvas fingerprinting, we exclude: (1) all the canvas with width and height below 16px; (2) scripts that do not use at least two colors or text with more than 10 different characters; (3) scripts that do not call either the toDataURL or the getImageData methods with an area below 320px; and (4) scripts that use the save, restore, or addEventListener methods of the rendering context. Despite these precautions, none of the scripts reported by OpenWPM meet these criteria. As a result, we set stricter conditions to identify scripts performing font fingerprinting: we only count those that set the font property and call the measureText method on the same text at least 50 times. This allows us to find 245 different JavaScripts performing canvas fingerprinting in 315 porn websites. 74% of the JavaScripts are fetched from 49 third-party services including ero-advertising.com and highwebmedia.com, a service that belongs to chaturbate.com (one of the biggest live sex services). These third-party services are present in 4% of all the porn websites in our dataset. We only find one script, delivered by online-metrix.net, using font fingerprinting.

We find that the script performing font fingerprinting and 91% of the scripts using canvas fingerprinting have not been previously indexed by tracking blocklists like EasyList and EasyPrivacy. As a result, these services could track users even when they use plugins such as ABP [102]. One example is the script delivered by xcvgdf.party (see Table 3.5) which performs canvas fingerprinting on

18 different porn websites, including a website offering transsexual/transgender pornography, `ladyboy-porno.com`.

3.3.1.4 *Other Potential Tracking Methods*

Our methodology allows us to find instances of other methods that could be potentially used for tracking purposes. However, we did not gather sufficient evidence to demonstrate that these JavaScript APIs are actually used for such purposes. One case is WebRTC [122], a technology to establish real-time peer-to-peer communications between browsers. WebRTC APIs allow collecting the IP address of the users, as well as the local network address. Through the combination of WebRTC with other tracking techniques [24], online services can discover networking information such as devices hosted behind the same NAT for cross-device tracking [123], or identify whether the user connects through a VPN [93]. In our dataset, there are 27 different JavaScripts using WebRTC present in 177 different pornographic sites, 21 of which use other tracking mechanisms in conjunction. Two of the 13 different third-party services using WebRTC, appear in the regular web and are classified as ATSes by EasyList. These services are `traffichunt.com` and `online-metrix.net`, an advertisement platform and a web analytic service, respectively.

3.3.2 *(Lack of) Network Security Standards*

Safeguarding users' privacy and security should be a priority for providers of pornographic content, particularly if users can be subject to censorship and surveillance at the network level [124, 125]. The use of encryption for transmitting data over the network is also a provision in privacy laws such as the GDPR (Article 32 [1]) and CCPA [2]. To identify the lack of security protocols in pornographic websites, we measure HTTPS support in porn websites by inspecting the requests triggered by our OpenWPM crawler. By default, we crawl each website using HTTPS, only downgrading to HTTP when HTTPS is not supported by the server.

Table 3.6 shows the use of HTTPS in pornographic websites depending of their highest Alexa rank in 2018. We find that over 92% of the most popular websites (in the top-1K of the Alexa ranking) do support HTTPS. However, the ratio of porn websites supporting HTTPS drops as their popularity does: HTTPS support decays to less than 25% for websites whose highest Alexa rank in 2018 was 10,000 or lower. This trend is similar for third-party services: those included in popular porn websites are more likely to support HTTPS. Nevertheless, we can find that 4,663 pornographic websites (68% of the total) are not fully HTTPS: either the website or one of its embedded third-party do not support HTTPS. By inspecting the content of these flows, we can identify that 8% of these websites upload cookies containing sensitive data in the clear as shown in Section 3.3.1.1.

Interval	Feature	HTTPS
0 — 1k	Porn websites (75)	92%
	3rd-party services (407)	90%
1k — 10k	Porn websites (552)	63%
	3rd-party services (1,327)	48%
10k — 100k	Porn websites (3,886)	32%
	3rd-party services (3,702)	25%
100k+	Porn websites (2,330)	22%
	3rd-party services (2,363)	16%

Table 3.6: HTTPS usage in pornographic websites

3.3.3 *Potential Malicious Behaviors*

We conclude this section with a short study of the presence of potentially malicious behaviors in pornographic websites according to VirusTotal [**vt**]. To minimize false positives, we only report domains flagged as malicious by at least 4 of the 70 different malware scanners aggregated by VirusTotal. There are 7 porn websites classified as a potentially malicious by VirusTotal. Further, malicious and deceptive behaviors also extend to 16 third-party services embedded in 41 porn websites.

We highlight the presence of three cryptocurrency mining services: `coinhive.com`, `jsecoin.com` and `bitcoin-pay.eu` in 8 porn websites. The latter domain, `bitcoin-pay.eu`, is not active anymore, but is related to `crypto-webminer.com` [126]. This suggests that owners of pornographic websites had explored alternative monetization schemes beyond online advertisement and subscription-based models. Whether these practices are performed with user consent is beyond the scope of this study.

3.4 MEASURING GEOGRAPHICAL DIFFERENCES

This section measures whether pornographic websites adapt their behavior – including the presence of trackers – to the geographical location of the user, possibly to meet the requirements of different regulatory frameworks. For that, we launch our crawls from different vantage points using commercial VPNs and our physical vantage point located in Spain.

3.4.1 *Third Party Services*

Table 3.7 shows the number of third-party services embedded in porn websites per country. We can see that the total number of third-parties in each location remains rather stable but for Russia, which has over 700 third-party services

	FQDN	Web Ecosystem	Unique Country	ATS	Unique Country
USA	5,483	16%	357	635	25
UK	5,364	15%	231	620	20
Spain	5,494	16%	561	592	59
Russia	4,750	16%	373	542	27
India	5,340	15%	275	607	21
Singapore	5,310	15%	233	608	16
Total	7,813	14%	2,030	816	168

Table 3.7: Comparison of the domains found on porn ecosystem from different geographical points. The values do not include domains loaded dynamically on the websites.

less. When looking at individual instances of third-party services, we find that there are hundreds of domains that are unique in each country but around 10% of them are related to CDNs or porn websites that generate arbitrary domains such as `img100-589.xvideos.com`.

If we look at ATS domains specifically, we can see that Google services dominate at a global scale, regardless of users' geolocation.

3.4.2 *Malware Presence*

The number of third-party domains considered as malicious by VirusTotal varies per country: from 15 third-party domains when accessed from Russia to 19 when accessed from India. Yet, 13 of these malicious domains are present regardless of users' geolocation (*e.g.,* the cryptomining domain `coinhive.com`). When counting the number of pornographic websites that contain such malicious content, the figure varies from 29 websites in Russia to 42 in Spain. Nevertheless, 26 pornographic websites always contain malware regardless of the country of access. This indicates, that some of the organizations serving malicious content might target users located at specific world regions.

3.5 REGULATORY COMPLIANCE

We now evaluate pornographic websites' efforts to comply with regulation. Specifically, we verify: (1) the presence and use of cookie consent forms as required by the EU GDPR and ePrivacy regulation; and (2) the use of verifiable age verification mechanisms in the context of the UK's Digital Economy Act. We also investigate the lack of privacy policies and potential inconsistencies that exist between these legal documents and the behavior observed in each pornographic websites in terms of user tracking and the presence of third-party ATSes.

Type	EU	USA
No Option	1.36%	1.39%
Confirmation	2.82%	2.3%
Binary	0.2%	0.06%
Others	0.03%	0.01%
Total (N = 6,843)	4.41%	3.76%

Figure 3.5: Usage of HTTP cookie banners in porn websites.

3.5.1 *Cookie Consent Notice*

The ePrivacy directive will require websites to obtain consent from European users before installing and using cookies, unless the cookie is strictly necessary for the webpage functionality. As this legislation is not yet into effect (it will take effect in 2019), the use of cookies is currently regulated by the GDPR, which indicates that users must consent to the use of any technique that may uniquely identify them [1]. This is typically done through cookie consent forms.

Degeling *et al.* performed a preliminary analysis of cookie consent forms (cookie banners) in 6,579 websites after GDPR came into effect [37]. They found that around 62% of the websites display a cookie consent-banner and developed a categorization of HTTP cookie banners that considers 6 different groups: (1) *No Option*: This type of cookie banner only informs users about the use of HTTP cookies without giving the possibility of accepting or rejecting them; (2) *Confirmation*: This type informs users about the use of cookies, but users can only show their accordance with the use of cookies, they can not reject them; (3) *Binary*:In this case, users can accept or reject the use of cookies; (4) *Slider*: This type of cookie banner gives users more fine-grained control over the level and type of cookies, that they allow by adjusting a slider; (5) *Checkbox*: This type of banner gives users the capacity to allow/reject cookies for a specific purpose or from a particular third-party service; and (6) *Other*: Any other type of banner that does not match any of the above. These banners tend to have a higher degree of complexity.

Identifying cookie banners automatically in websites following Degeling's method and taxonomy is not trivial. In fact, we could only instrument our customized OpenWPM to identify the following types: *No option*, *Confirmation* and *Binary*. We merge the *Slider* and *Checkbox* types together in the *Others* category, as we would need to interact with the banner to be able to further categorize them. Our method works as follows: first, we inspect the HTML DOM to find elements that resemble a banner (inspecting the text of the banner). If such an element is found, we extract the text rendered to the user, and take a screenshot that we manually analyze to manually verify that the HTML element is indeed a banner. We repeat this procedure from two countries, Spain and the USA to find potential differences in cookie banner presence.

Table 3.5 shows the percentage of pornographic websites in which we find HTTP cookie banners. As can be observed, the proportion of pornographic webpages with cookie banners is very small, being only 4% of the total. A second observation from Table 3.5 is that the difference between accessing webpages from one country or the other is also very small, as only 0.65% more pages show a cookie banner when fetching them from Europe.

The low presence of cookie banners is remarkable when compared with the fact that 72% of the pornographic websites studied contain third-party cookies (Section 3.3.1.1).[11] Moreover, out of the websites that show a cookie banner, 32% do not give users any control over the use of cookies as the banner only discloses their use (No Option type). While it is possible that not all third-party cookies are actively used for tracking purposes, these figures suggest that many websites offering sensitive content may potentially be in violation of the GDPR.

It is important to note that our methodology uses OpenWPM to crawl the websites and that we do not interact with the webpage once we have visited it. Therefore, even in the websites where a cookie banner is present, we never gave actual consent to the use of cookies.

As a final note, one might expect that large corporations providing pornographic content would have strong incentives to be in compliance with regulatory requirements. While this is the case for a small fraction of popular pornographic websites, there is not a clear correlation between the use of cookie consent forms and the popularity of porn webpages.

3.5.2 *Age Verification*

Some pornographic websites have taken positive steps to implement age verification mechanisms in an effort to comply with increasing regulatory pressures (see Section 2.2.2). In this section, we study how prevalent and how effective verifiable age-verification mechanisms are in the wild. For that, we use our Selenium-based crawler to parse the landing page of each porn website, and look for warnings and consent forms displayed to the user as detailed in Section 3.1.1. As our approach relies on string matching to identify such warnings, it is prone to introduce false positives, specially so in age-related keywords that appear often in the content of the websites. Therefore, due to the difficulty to perform this study automatically and at scale, we only investigate a subset of the top-50 most popular pornographic websites manually.

We perform this manual analysis in 4 countries (the US, the UK, Spain, and Russia) to identify regional differences. The results from the USA, UK and Spain are consistent: *i.e.*, the same set of 20% of the pornographic websites implement and show to the end user the same age verification mechanism, consisting of a simple warning text and a button to be clicked on. However, there are significant differences when accessing the same websites from Russia: only 14% of the analyzed websites have an age verification mechanism. Additionally, 8% of the websites that do not verify users' age for the rest of countries do so in Russia,

11 For comparison purposes, Degeling *et al.* showed that 69.9% out of a corpus of 6,357 websites had a cookie consent banner in January 2018 [37].

whereas 12% of the websites do not verify user age in Russia but do so in the rest of countries studied. We note that we did not find any instance of AgeID being deployed during our study.

Despite regulatory pressures, the current age verification mechanisms implemented by all these sites are easy to bypass and could not be considered as "verifiable age verification mechanisms". In other words, if our automatic crawler manages to bypass the mechanism, a child could do it as well. We only found one webpage in Russia, `pornhub.com`, implementing a complex age verification mechanisms through social media accounts as requested by the Russian federal government in 2017 [62].

3.5.3 *Privacy Policies vs. Reality*

The GDPR [1] requires all websites collecting or processing personal identifiable data from European citizens to portray a privacy policy describing their personal data collection and processing practices, including data collected by embedded third parties. We perform a best-effort crawl to collect the privacy policies, if available, of each pornographic website to crosscheck with our empirical results, and highlight potential privacy violations. We perform this analysis using the method introduced in Section 3.1.1, only from our physical machine located in Spain.

Our crawler inspects the DOM of the landing page looking for a link to the privacy policy. We are able to find a privacy policy in 16% of the pornographic websites in our dataset. We get these figures after a manual sanitization of our results in which we manually check the privacy policies which are abnormally short in the number of words and found 44 false positives caused by HTTP errors (response codes).

The GDPR forced changes in the way privacy policies are presented to users, forcing publishers to be clear about their data collection, processing and sharing practices, as well as user rights. We use string matching to find that 218 (20%) of the privacy policies make an explicit mention to the GDPR. We dive deeper into the analysis of the privacy policies by first looking at length patterns, in an attempt to understand how similar (or different) policies are. We find that, on average, privacy policies contain 17,159 letters and that the shortest policy we found has 1,088 letters, and the largest 243,649.

While this might hint that there are big differences across policies, we further investigate the similarity of the text in privacy policies. We use the term frequency-inverse document frequency (TF-IDF) [96] to measure the similarity between two texts.[12] We run this measure for the 1,202,312 pairs of different privacy policies in our dataset and found that 76% have a similarity above 0.5 (meaning they are co-related). This can be a direct result of websites belonging to the same company having a very similar privacy policy as well as the prevalence of templates that are highly popular across websites. In fact, finding pairs of websites with a coefficient of 1 helped us discover companies holding a larger number of pornographic websites (Section 3.2.1).

12 The value goes from -1 (exactly opposite) to 1 (exactly equal) going through 0 (no co-relation).

The opacity of the privacy policies makes it difficult to perform an automatic analysis of their content at scale. To tackle this issue, we use the publicly available tool Polisis[127], which presents a human-readable summarized version of the privacy policy, to extract third-party entities and data collection methods. As Polisis does not provide APIs to access the results in a machine-readable format, we rely on the web version of the tool to further investigate the top 25 websites tracking users (*i.e.*, canvas fingerprinting and cookies) according to our results from Section 3.3. We manually asses that 72% of this subset of porn websites have a privacy policy in which they clearly state the use of cookies, the type of data collected, and the presence of third parties in their websites. Only one of the websites discloses in its privacy policy the complete list of third-party advertising and tracking services.

These results show that – while privacy policies are becoming more common, complete, and clear to users – there are still many websites engaging in user tracking without privacy policies and other transparency mechanisms. When they do, with only one exception, they do not disclose the whole list of embedded third-parties.

Part III

AN ANALYSIS OF DOMAIN CLASSIFICATION SERVICES

4

AN ANALYSIS OF DOMAIN CLASSIFICATION SERVICES

The need to classify websites became apparent in the early days of the Web. The first generation of domain classification services appeared in the late 1990s in the form of web directories. Notable examples from this period are Yahoo! Directory [128] and DMOZ[1] [130]. The main purpose of such services was to facilitate the discovery of web pages relevant to a certain topic of interest. To this end, human editors manually classified sites—often relying on suggested categories submitted by other users—into a purpose-specific taxonomy [131]. The quick expansion of the Internet soon put this approach to an end and led to the development of automated classification solutions [132, 133, 134, 135, 95].

As the Web grew in size, content, and applications, domain classification services became a valuable facilitator in multiple areas. One key application is traffic filtering, *i.e.,* networking solutions designed to block access to sites that are deemed dangerous (*e.g.,* phishing or malware [136, 137]) or inappropriate (*e.g.,* adult content). Cybersecurity firms such as McAfee [138] and OpenDNS [139] (Cisco) rapidly developed their own products. These technologies are nowadays embedded in multiple applications and setups such as parental control solutions and traffic filters in schools [140], libraries, and enterprise networks [141, 142, 143]. The online marketing industry also found domain classification extremely useful, in particular to improve targeted contextual advertising [144, 145, 146]. This led the Interactive Advertising Bureau (IAB) to develop an open standardized taxonomy for real-time bidding protocols [147]. Finally, networking, privacy, and security researchers also rely on website classification services to conduct category-dependent measurements [148, 149, 150] or to discover websites falling in a given category [151, 152].

However, no study so far has specifically analyzed the coverage, labels and applicability of domain classification services in different scenarios and research domains. Classifiers that were developed for different target applications or with different methodological approaches often exhibit disparate characteristics in terms of their coverage and taxonomies. This may have a substantial impact on how much the applications and studies that rely on them can be trusted. In fact, previous research studies reported the need for manual classification of websites due to the shortcomings of commercial services [36, 95, 153]. Unfortunately, the evaluation of these services is complicated by their opacity. While many services claim to apply machine learning algorithms, it is unclear how thoroughly they perform concrete analyses to validate their solutions, how comprehensive the underlying training data is, and, ultimately, how trustworthy and accurate the resulting classification is. Similarly, services such as DMOZ and OpenDNS that rely on human volunteers may be biased due to subjective

1 DMOZ was closed in 2017 by its operator AOL. It has been continued by the Curlie project, which is still operating as of this writing [129]

opinions in the moderation process. Therefore, classification services may not succeed at adequately covering the large diversity of websites in both number and nature.

We address the questions above by presenting a first analysis of domain classification services. Specifically, we make the following contributions:

1. We analyze 13 popular services selected through purpose-specific web searches as well as through a survey of all the academic works published during 2019. We find that the results of 24 academic papers published in 9 relevant conferences (*e.g.,* IMC, WWW) depend on the outcome of the domain classification services that they use. Then, we present a qualitative analysis of the approach followed by these domain classification services according to their documentation. We find that key differences in their approaches might affect coverage and accuracy.

2. We evaluate the coverage of these services for both popular and unpopular domains, their labeling methodology, their taxonomies, and the agreements across services when labeling the same domains. We crawl the labels assigned to 4.4M domains and find that most services lack coverage (only two services have a coverage above 55%), especially for non-popular domains. Furthermore, we show that their complex taxonomies (in particular for marketing-oriented classification services, with sometimes over 7.5k observed labels) hinder sound interpretation.

3. We study how introducing humans in the labeling process might impact the coverage and label consistency of those services. We find that manual classification is affected by disagreements, ambiguities, and mismatches in the labeling process as well as biases in the distribution of users that submit votes and the workload of editors. This translates in some domains receiving as many as 58 rejected labels. To gain a better understanding of these challenges, we run a controlled experiment involving manual domain labeling and find disagreements in 35.5% of the cases.

4. We explore the performance of domain classification services as tools to identify websites of interest. To do so, we run three case studies in the areas of detecting (and filtering) advertisement and tracking, adult content, and CDN or hosting infrastructure. We find that the accuracy and coverage of the studied services is extremely low, and that the choice of one service or another significantly affects the outcome because of differences in coverage, which ranges from over 95% to below 1%.

5. Finally, we discuss the implications of our findings for both the technical and academic applications of these services. We also provide recommendations on how users should handle the significant disparities observed across services and identify a number of research questions for future work.

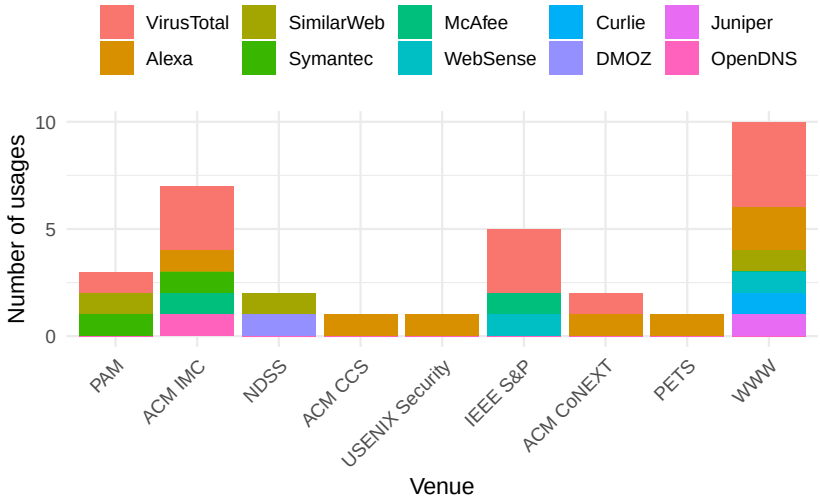


Figure 4.1: Usage of domain classification services in research during 2019. We have not observed the use of these services in TMA and ACM SIGCOMM papers in 2019.

4.1 USAGE IN ACADEMIC STUDIES

Domain classification services play a fundamental role in academic studies. Given that the unknown properties of these services can impact research results, it is important to understand how widespread their usage is and what they are used for in the literature.

Survey approach. We survey all 1,014 papers published in 2019 at top venues in four areas: *i)* network measurements (IMC, PAM, TMA, CoNEXT, SIGCOMM); *ii)* security and privacy (CCS, NDSS, S&P, USENIX Security, PETS); and *iii)* Web (WWW). We first search for the names of domain classification services as well as keywords that indicate that such a service is used.[2] We then discard obvious false positives, such as the Amazon Alexa voice assistant instead of the Alexa domain classification service.

Usage. We manually analyze the remaining papers and find 26 papers that use at least one domain classification service (Figure 4.1). We find that for 24 (92%) of these, their results depend on the choice of service as they use it to gather their initial dataset or validate their results. Papers accepted at WWW and IMC are the ones that tend to rely the most on domain classification services. VirusTotal is the most popular service among academic studies (12 papers). Specifically, 3 papers [154, 155, 156] use the aggregate of VirusTotal's categories while 3 others [157, 151, 150] select one or more of the specific providers integrated in this popular threat-intelligence service. The remaining 6 papers [158, 159, 160, 161, 162, 163] only rely on VirusTotal's detection of malicious domains or files. The second most popular service is Alexa, with 7 papers relying on it. All of these papers use Alexa's lists of top sites per category to gather a corpus of websites (*e.g.*, governmental [164] or gambling and dating websites [151]). One paper [41]

2 The keywords used are "website classification", "website categorization", "domain categorization", "categorization service", "website category", "domain category", "category of the website", and "category of the domain", in singular and in plural, and also using British English spelling.

also uses the list of top sites per country. Our analysis reveals one paper using SurfControl [165], but as this service was acquired by Websense in 2007 [166], we do not consider it further. Table 4.1 lists all analyzed publications per venue.
Purpose. The 26 papers using domain classification services do so for a wide range of purposes. We find that 9 (35%) of them focus on security topics, including mobile sensors attacks [163] and certifications in the online payment industry [167]. We find 4 (15%) papers studying privacy in specific website categories—*e.g.*, tracking on pornographic websites [168] —or email tracking [155]. We identify 6 (23%) measurement papers, *e.g.*, on resource reloading by third-party websites [150] or web complaints [156]. Finally, 4 papers *question* the accuracy and applicability of existing domain classification services and either choose not to rely on them [153, 36] or manually validate the results [37, 169].

Venue	Area	Papers	using service		# dependent	References
			#	%	on service used	
TMA	Measurements	24	0	0%	—	—
PAM	Measurements	20	3	7%	3	[148, 161, 170]
ACM IMC	Measurements	39	5	12%	5	[168, 157, 158, 171, 154]
NDSS	Security	90	1	1%	1	[172]
ACM CCS	Security	148	1	0.7%	1	[167]
USENIX Security	Security	112	1	0.9%	0	[169]
IEEE S&P	Security	90	4	4%	3	[159, 160, 173, 155]
PETS	Privacy	68	1	1%	1	[41]
ACM SIGCOMM	Networking / Systems	31	0	0%	—	—
ACM CoNEXT	Networking / Systems	32	1	3%	1	[151]
WWW	Web Tech.	360	9	3%	9	[152, 149, 174, 162, 164, 150, 163, 156, 175]
Total		**1,014**	**26**	**3%**	**24 (92%)**	

Table 4.1: Usage of domains classification services in the literature in 2019. The "dependent" column indicates whether the results of the study depend on the quality of the service used.

Takeaway: *We find that 26 papers published at top peer-reviewed conferences from 2019 use domain classification services. For 92% of these, their results depend on the choice of service, even though these services are sometimes questioned. As we will show later, in the absence of ground truth this dependence can introduce biases in the study results.*

4.2 PROVIDER ANALYSIS

We examine the claims made by classification services (if available) in terms of their purpose, methods used for classification, coverage of URLs and languages, and development of their taxonomy. We retrieve these details through a manual inspection of their own documentation.
OpenDNS. OpenDNS provides DNS-based content filtering, sourcing website categorization from its human volunteer-based "Domain Tagging" project [139]. Participants submit domains and their categories, on which other participants may vote; once the mapping of a domain to a category receives sufficient votes, it is available for approval by a community moderator before it is propagated

to the content filtering system [176]. These moderators also review reports of incorrect categorization as well as categories of popular sites [177]. We expand on the effects of this voting procedure in Section 4.5. OpenDNS has at least one confirmed category for almost 4 million domains, out of 12.7 million submitted domains [139]. A list of categories and short descriptions is available [178]. Users had the ability to suggest the addition of categories to the taxonomy [177]; it is unclear who approved these new categories.

McAfee. McAfee provides the "TrustedSource" online service (previously called "SmartFilter") for obtaining both the category and a reputation score-based risk assessment for a URL [138], mainly with the goal of client-side content filtering. A user of the service must choose one of eight 'products', which affects the 'URL Filter database' version used. Categories are specific to URLs. McAfee categorizes web pages through "various technologies", including both machine learning and manual review [179]. It is said to cover "millions of Internet sites" [179]. McAfee's category taxonomy is documented in detail, listing descriptions, examples and related categories as well as taxonomy updates [179]. However, this document was last updated in 2010.

FortiGuard. FortiGuard provides an online tool for retrieving content-based URL categorization [180], which supports the content filtering functionality in its FortiOS-based FortiGate firewall [181]. Websites are classified through a "combination of proprietary methods including text analysis, exploitation of the web structure, and human raters" [181]. FortiGuard's service is said to include over 45 million website ratings that cover over two billion URLs [181]. Categories are divided into seven high-level groups (adult, bandwidth-consuming, business, personal, potentially liable, security, and unrated), and short descriptions and test pages are available [182].

VirusTotal. VirusTotal is an online service providing analysis of potentially malicious files and URLs by aggregating the results from a large set of detection engines [183, 184]. It also lists the domain's category, but it is unique among the other services in that it does not establish its own categorization. Instead, it collects labels from existing services: at the time of our data collection, these were Alexa, Bitdefender, Dr.Web, Forcepoint, Trend Micro, and Websense, but since July 2020, these were (at least) Bitdefender, Comodo Valkyrie Verdict, Dr.Web, Forcepoint ThreatSeeker, Sophos, and Yandex Safebrowsing. For each service, VirusTotal displays at most one distinct label, without combining labels any further, *i.e.,* a domain can have as many categories as there are services. Categories are only provided for domains, even though a user can also request scanning for URLs.

Alexa. Alexa offers the ability to view the 500 most popular websites for a specific category [185], with a focus on marketing and content discovery. Its results are based on the human volunteer-based categorization from DMOZ [186], but in contrast to DMOZ, Alexa's lists only contain domains, not URLs. Alexa's taxonomy is also based on the DMOZ's taxonomy, but pruned to around 280,000 categories. Alexa does not allow searching for the category of a specific domain. The ranking within a category is calculated using the same methodology as the main Alexa top list, but if applicable only using the data for the specific subdo-

main [187]. As the main Alexa top list only lists base domains, this may result in a different relative ranking for two domains [187].

Bitdefender. Bitdefender provides content category-based website filtering in its consumer- and business-oriented products [136]. There is no free online categorization tool, but VirusTotal integrates Bitdefender's categorization into its domain analysis. Its database is said to cover millions of URLs in multiple languages [188]. A list of (ungrouped) categories, short descriptions, and examples is available [189].

Forcepoint/Websense. Forcepoint (renamed from Websense in 2016 [190]) provides an online tool for website threat and content analysis [191]. The tool shows both a static (*i.e.,* previously determined) and a real-time classification. The former results from a combination of automated and manual inspection [192], while the latter is based purely on an automated machine learning-based approach [191]. Forcepoint will classify the specific page of a given URL, not its base domain [193]. Categories are divided into six high-level groups (reputation, security, bandwidth-consuming, productivity-inhibiting, social networks, and baseline) for which short descriptions are available [192].

Dr.Web. Dr.Web includes a category-based website filter in its client-side antivirus software, but its online tool only provides a binary classification of a URL's maliciousness [194]. A more detailed categorization is accessible through VirusTotal, but appears to only cover types of malicious behavior. No documentation is available on the categorization process or the possible categories.

Trend Micro. Trend Micro's classification security-oriented service is available online through its "Site Safety Center" [195]. Next to a content-based category, they establish a threat rating denoting whether a website is 'safe', 'dangerous', 'suspicious' or 'untested' [195].Their database is said to include over 35 million URLs, and they acknowledge that "a few URL rating errors" may occur [196]. Trend Micro publishes two lists of available categories with short descriptions. One was last updated in late 2019 and appears to be used for their "Worry-Free Business Security" and "OfficeScan" web threat protection products [197]; its categories are grouped into seven 'filtering groups'. The other was published at the latest in November 2011 [198] and has not been updated since [199]; its categories are ungrouped.

Symantec. Symantec (now part of Broadcom) provides an online tool to retrieve the URL categorization from its WebPulse system [137], which powers its web gateway content filtering. The categorization system is said to use manual and automated (machine learning) analysis, with several modules voting towards the final categorization [200, 201]. The tool indicates how recently the URL was categorized; previously unknown URLs are purported to be classified in real time [200]. Its URL database is said to cover "millions of entries", and supports over 60 languages [201]. A URL can be classified as up to four categories [200]. A listing of categories, descriptions, examples and test sites is available [202]. The taxonomy was last updated in August 2019 [203].

Webshrinker. Webshrinker provides an online demo tool of their URL categorization service [145]. Their service targets two audiences: a purely content-based categorization aimed at advertisers, and a security-oriented service which

combines custom heuristics, machine learning, internal and external data feeds to assess web threats [145, 204]. Classification is said to occur in real-time [205], their database covers over 97.2 million 'entries' [204], and they support over 12 languages [205]. The two target audiences are also reflected in the two available taxonomies [205]. One is a custom list of 42 'standard' categories designed for content filtering, while the other uses the taxonomy of over 390 categories developed for marketing purposes by the Interactive Advertising Bureau (IAB) [147]. For the latter, Webshrinker computes a confidence score [206].

DMOZ. DMOZ (also known as the Open Directory Project) operated a directory of web pages, where users could navigate the category structure to find URLs in that category [130]. Its owner AOL took down DMOZ in 2017 after 19 years of operation [207]. DMOZ's rich taxonomy consisted of sixteen top-level categories, each being the top leaf in a large hierarchy of gradually more fine-grained subcategories, amounting to over a million categories encompassing 3.86 million URLs [130]. All users could suggest the addition of a URL to a category, but this had to be approved by one of the 91,929 category-specific editors [208]. Editors were also responsible for developing subcategories of the categories they maintained, which was suggested they do once a category reached 20 links [209]. DMOZ had strong multilingual support, with separate directories for 90 languages [130]. DMOZ allowed to search whether and where URLs appear in the directory.

Curlie. The Curlie project [129] emerged as the successor of DMOZ. Curlie retains the community of human editors, who appear to continue updating the directory listings and taxonomy to this day. Compared to its predecessor, Curlie has around 500 more categories (out of 1 million), but around 500,000 fewer URLs, and support for two more languages [129]. Like its predecessor, Curlie allowed to whether and where URLs appear search in the directory.

4.3 METHODOLOGY OF DOMAIN CLASSIFICATION SERVICES

We perform an analysis of the 13 domain classification services listed in Table 4.2 using publicly available information. We select them based on their usage in recent academic works (Section4.1), extending the set with services found through targeted online searches. Note that 2 of the domain classification services that we consider (FortiGuard and Webshrinker) were not used by any of the surveyed academic papers published in 2019. Our list does not cover all commercially available services, but those omitted pose a high barrier for data collection because of technical or monetary reasons.[3]

Furthermore, VirusTotal is unique in that it does not provide its own classification, but instead aggregates category labels from third-party scanners. At the time of our data collection, these scanners were Alexa, Bitdefender, Dr.Web, Forcepoint, Trend Micro, and Websense.[4] However, since July 2020, these consist of (at least) Bitdefender, Comodo Valkyrie Verdict, Dr.Web, Forcepoint Threat-

3 *e.g.*, Zvelo and Cyrenrequire completing a reCAPTCHA for every request.

4 Websense renamed itself to Forcepoint [190] after the acquisition of Stonesoft, yet both are listed separately in VirusTotal.

		Input		Output	Purpose				Updates			Access	
Service		Domain	Subdomain	Multiple categories	Content filtering	Threat assessment	Marketing	Discovery	Automated	Real-time	Reclassification	Free (sample)	Documentation
OpenDNS	[139]	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	[178]
McAfee	[138]	✓	✓	✓	✓	✓	✗	✗	✓✗	✗	✓	✓	[179]
FortiGuard	[180]	✓	✓	✗	✓	✗	✗	✗	✓✗	✗	✓	✓	[182]
VirusTotal	[183]	✓	✓	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
Alexa	[185]	✓✗	✓✗	✓	✗	✗	✓	✓	✗	✗	✓	✓	[186]
Bitdefender	[136]	–	–	–	✓	–	✗	✗	✓	–	✗	✗	[189]
Forcepoint	[191]	✓	✓	✗	✓	✓	✗	✗	✓✗	✓	✓	✓	[192]
Dr.Web	[111]	–	–	–	✓	✓	✗	✗	✓	–	✗	✗	✗
Trend Micro	[195]	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	[197]
Symantec	[137]	✓	✓	✓	✓	✗	✗	✗	✓✗	✓	✓	✓	[202]
Webshrinker	[145]	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✓	[205]
DMOZ	[130]	✓	✓	✓	✗	✗	✗	✓	✗	✗	✓	✓	[130]
Curlie	[129]	✓	✓	✓	✗	✗	✗	✓	✗	✗	✓	✓	[129]

Table 4.2: Features of the analyzed classification services. For the services aggregated in VirusTotal, we list their properties as if they were accessed directly.

Seeker, Sophos, and Yandex Safebrowsing. We consider the former services (independently) in our evaluation. In Section 4.4.2, we evaluate the consistency of services across multiple available sources.

Our evaluation focuses on features and methodological aspects that might affect how these services can be used in technical solutions and academic studies. Table 4.2 shows the features exhibited by the selected services according to their documentation and websites. We also register our own domain and set up a live website hosting a WordPress blog, and then request its classification from each provider to investigate their approach to classifying new domains. We consider the following properties:

Inputs. The granularity of input provided to the classifier affects the correctness of the classification: a subdomain may host a different kind of content than its base domain. For example, subdomains of the base domain (`yahoo.com`) may host a search engine (`search.yahoo.com`), a sports news site (`sports.yahoo.com`), or a webmail service (`mail.yahoo.com`). Depending on the origin of domains to be classified, *e.g.*, domain top lists often used by researchers that can include subdomains [210, 211], this can impact the accuracy and perception of the labels. All evaluated services may provide a separate classification for a subdomain. However, Alexa does not have a way to retrieve the classification given

a (sub)domain. Instead, it requires searching through its listings of the top 500 domains in one of 279,716 categories.

Outputs. The outputs affect the utility of the data to a study's purpose. If a service yields multiple categories for a given site, this may improve the applicability and correctness of the classification as it can be more nuanced, *e.g.*, tagging a sports news website as both *sports* and *news*. However, this could also lead to an incoherent interpretation, *e.g.*, double-counting when aggregating domains by category. All services except FortiGuard and Forcepoint can assign multiple categories to domains.

Purpose. In many cases, the provider's intended purpose for a service (*e.g.*, content filtering, threat protection, marketing or discovery of relevant content) influences the used taxonomy. For example, a content-filtering service may prefer to label `youtube.com` purely as a bandwidth-consuming site, but a marketing-oriented service may label it as a video sharing or advertising platform. Most of the classification services analyzed are intended for content filtering, usually being integrated into their consumer or business web security software. One exception is VirusTotal, which provides only a threat assessment. Further exceptions are Alexa, DMOZ, and Curlie, which are designed for discovering sites within categories of interest. Moreover, certain services also have other applications. For instance, Webshrinker can categorize domains according to the marketing-oriented taxonomy of the Interactive Advertising Bureau (IAB) [147].

Updates. The ability to update classification results affects both coverage and accuracy. Real-time classification, often enabled by a fully automated analysis, may improve coverage and maintain data relevance. In other words, new sites can be immediately assigned to a category, and the classification will reflect the most recent content. For example, a change in website ownership would not result in outdated labels. Automated approaches may also increase the scale at which domains can be classified, in particular when additional data is used to label uncrawlable domains (*e.g.*, malware domains). The ability to request reclassification of a site may allow to correct errors, but it may also be leveraged to undeservedly receive a less "harmful" classification if requests are not adequately reviewed. For example, an adult website may attempt to get reclassified as a (non-adult) video streaming site in order to evade filtering.

Only Forcepoint, Symantec and Webshrinker provide real-time results: we confirm through web server logs that upon request, they immediately visit and categorize a domain that we newly registered. Webshrinker even proactively visits the domain (likely due to its entry in the zone file), and is the only one to deploy a real browser. This behavior can be traced back to the methods that services claim to use, mostly consisting of automated classification through machine learning algorithms. McAfee [179], FortiGuard [181], Bitdefender [188], Forcepoint [192], Symantec [200, 201], and Webshrinker [204] state in their documentation that they complement their crawler-based ML solution with domain metadata, security honeypots and scanners, and third-party feeds and logs, as well as human reviewers who inspect and amend automatically determined categories. OpenDNS, DMOZ, and Curlie rely on human volunteers to propose and confirm categories; Alexa uses a truncated version of DMOZ's data and tax-

onomy [186]. All services except VirusTotal, Bitdefender and Dr.Web provide a way to request domain reclassification: for our newly registered domain, the delay of several days before any change suggests that this process requires human intervention.

Access. Easy access to data and documentation improves usability for end users and researchers. For instance, clear descriptions and examples of sites that are considered part of a category aid in selecting the appropriate categories for other websites. Bitdefender and Dr.Web do not provide direct free access to their data, but they are available through VirusTotal. Dr.Web is the only service that does not document its taxonomy. VirusTotal does not document where and how it sources its data. In Section 4.4.3, we compare the documented categories with those that we observe empirically.

Takeaway: *The substantial differences in domain classification services' characteristics affect their applicability: label interpretation depends on a service's supported inputs and outputs as well as taxonomy differences due to their purpose, while coverage and accuracy benefit from easy access to up-to-date labels. These properties should therefore be well understood to ensure correct application. We assess the veracity of services' claims through our own empirical observations in Section 4.4, to determine their effective suitability to different scenarios.*

4.4 DOMAIN LABELING QUALITY

In this section we analyze domain classification services on their labeling coverage (Section 4.4.2), their individual taxonomies (Section 4.4.3), and the labeling consistency and relationships across providers (Section 4.4.4). In this analysis, we omit DMOZ and Curlie as they aspire to achieve a different goal, *i.e.*, supporting content discovery instead of concisely classifying all domains. This affects their data retrieval strategy and interpretation, and we would need to reverse their mapping of deeply nested categories to relevant domains.

4.4.1 *Data Collection*

Our data collection process consists of two stages:

(1) **Compiling target domains.** We compile a large list of domains starting from the union of all daily Alexa top sites rankings between September 1 and 30, 2019. To reduce possible biases caused by the instability of the Alexa ranking [210, 211, 148], we aggregate these rankings using the default method of the Tranco top list [211], which sums domain scores from individual lists following a Zipf-like distribution. We retain a ranked list of 4,424,142 domains that we could successfully collect from all non-rate-limited services. While these 4.4M domains represent a small fraction of all registered domains [212], they are considered to be popular by the Alexa traffic ranking service. Their popularity is further reflected by the fact that 47% of the 4.4M domains are indexed in the Chrome User Experience Report [213] and 0.5% by Common Crawl [214], both generated between August and October 2019. We therefore believe that our set is represen-

tative of domains regularly visited by end users and therefore also of interest to researchers.

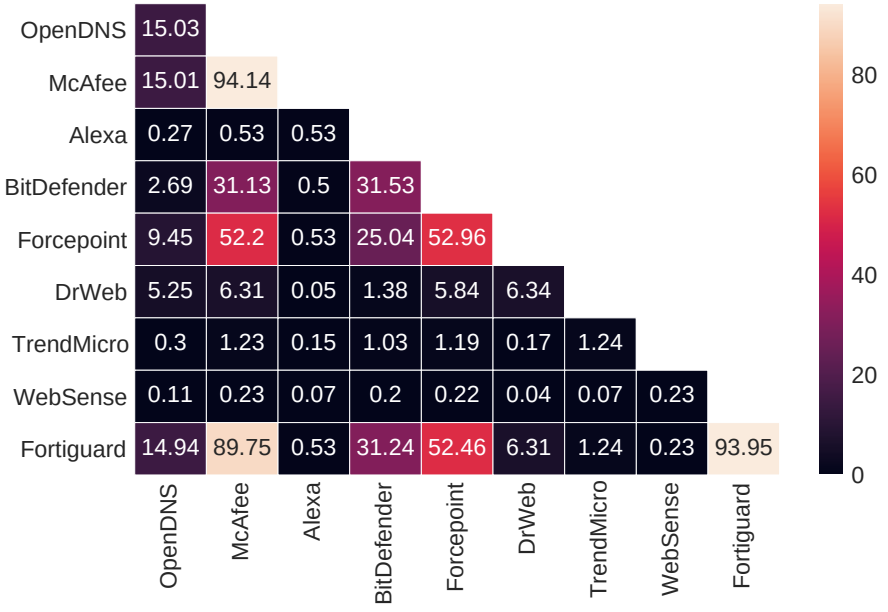
(2) **Crawling domain classification services.** We retrieve the category labels for the 11 selected domain classification services. As each service differs in how its online portals retrieve data, we develop the most scalable and least resource-intensive method possible for each provider.

- For FortiGuard, McAfee, and OpenDNS, we retrieve labels through their publicly available portals. While these services are not rate-limited and their data is public, we perform our data collection at a non-intensive average rate of 40 requests per minute. We retrieve McAfee's labels for its "Real-Time Database" product. For VirusTotal, we retrieve labels through its API, which aggregates six services: Alexa, Bitdefender, Dr.Web, Forcepoint, Trend Micro, and Websense. We received access to VirusTotal's academic API, with a request limit of 20k queries per day and account.
- For Symantec, Trend Micro and Webshrinker, our data collection is subject to rate limiting. Therefore, we retrieve labels on these three services for the top-10k domains in our ranked list. We retrieve Webshrinker's labels from its default marketing-oriented IAB taxonomy.

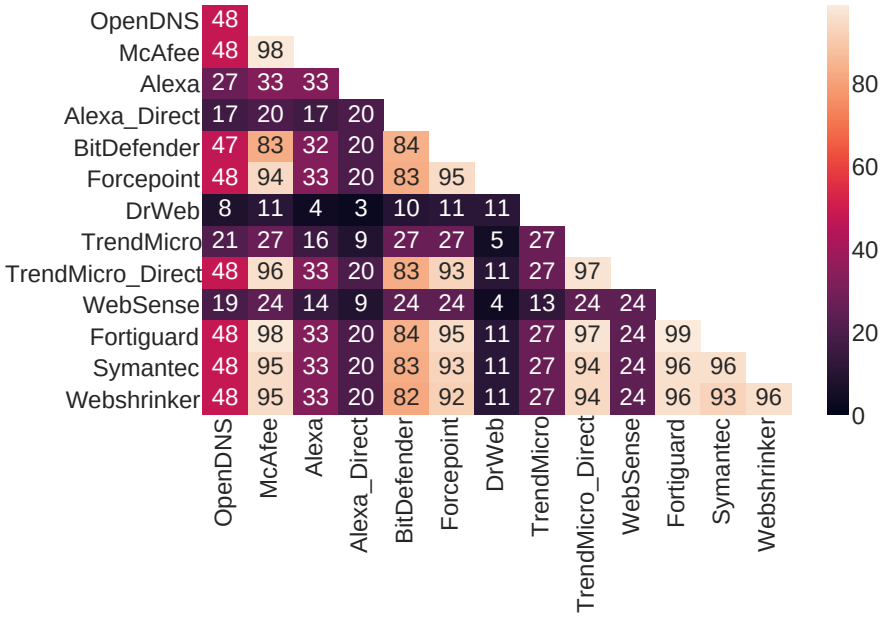
4.4.2 *Coverage*

One critical aspect to consider when using domain classification services is their coverage, defined as the number of websites for which they provide a meaningful label. This metric affects how comprehensively a service can both execute its original task and be deployed for large-scale applications and studies. As discussed in Section 4.3, some domain classification services involve humans in the loop, while others try to achieve a larger scale or real-time classification using machine learning methods. As a result, not all services have the same ability to scale their labeling process. When measuring coverage, we apply a sanitization process to address the fact that five services (FortiGuard, OpenDNS, Websense, Forcepoint and Trend Micro) provide explicit labels for unclassified domains. We consider a domain "unlabeled" if we obtain an empty result, or a label explicitly stating that the service has not (yet) labeled the domain (e.g. *Uncategorized* for Forcepoint).

Figure 4.2a shows for which percentage of our full set of 4.4M domains we obtain a valid label. The diagonal reveals that the coverage varies greatly between individual services. The off-diagonal values report the 'intersection coverage' defined as the number of domains that both services label simultaneously, regardless of the label provided. FortiGuard and McAfee excel by labeling around 94% of domains, likely due to their deployment of machine learning techniques for automated classification. Contrarily, OpenDNS only achieves 15% coverage, with its manual submission and voting processes (Section 4.5) likely becoming a bottleneck when dealing with the millions of monthly domain registrations [212]. Alexa's coverage is even lower at 0.53%, possibly due to its data source DMOZ [186] containing human-volunteered labels in often highly spe-



(a) Intersection (4,424,142 domains)

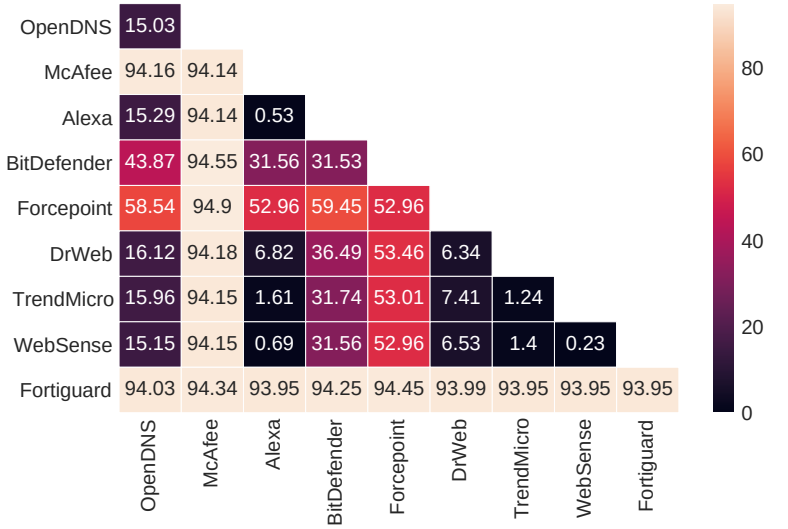


(b) Intersection (Top-10k domains)

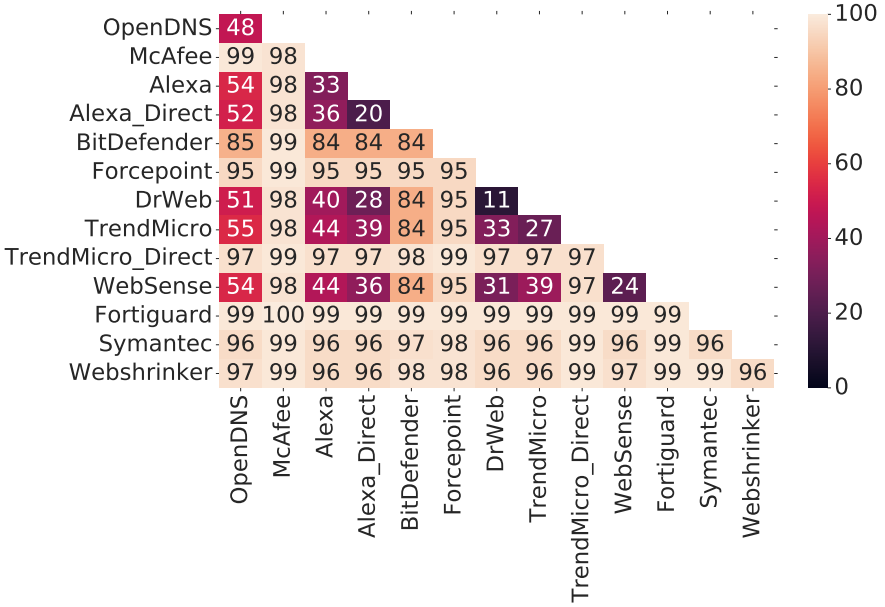
Figure 4.2: Coverage per service (diagonal) and intersection of the coverage between pairs of services for our two domain sets (Section 4.4.1).

cialized (and therefore less popular) categories designed for content discovery, as well as its limit of 500 websites per category. Services retrieved through VirusTotal also have much lower coverage; we will show later on that this may in part reflect a service integration issue t VirusTotal, as services do yield a label when directly queried.

For completeness, we also compute the "union coverage" between pairs of providers. We define it as the percentage of websites for which at least one service provides a valid label as we can see in Figure 4.3. This analysis suggests



(a) Intersection (4,424,142domains)



(b) Intersection (Top-10k domains)

Figure 4.3: Coverage per service (diagonal) and the union of the coverage between pairs of services for our two domain sets (Section 4.4.1).

that considering the union of two services does not necessarily increase the global coverage when their intersection is already high. For example, the union coverage for FortiGuard and McAfee increases slightly to over 98%. However, as we will discuss in Section 4.7, the combination of labels from multiple services is non-trivial due to largely disjoint taxonomies. As a result, unless the objective of unifying providers is offering complementary perspectives, it might not necessarily benefit coverage.

Rank	(0-1k]	(1k-10k]	(10k-100k]	(100k-1M]	+1M	Overall
# domains	1,000	8,945	89,276	678,246	3,646,675	4,424,142
OpenDNS	78%	46%	19%	9%	16%	15%
McAfee	100%	99%	98%	97%	94%	94%
FortiGuard	100%	100%	99%	97%	93%	94%
Alexa						
through VT*	48%	32%	13%	1.02%	0.05%	0.5%
direct source	31%	20%	-	-	-	-
Bitdefender*	93%	83%	73%	48%	27%	32%
Forcepoint*	98%	95%	90%	73%	48%	53%
Dr.Web*	16%	11%	6%	4.2%	7%	6%
Trend Micro						
through VT*	55%	25%	9%	2.7%	0.7%	1.2%
direct source	98%	97%	-	-	-	-
Websense*	52%	22%	3.9%	0.36%	0.04%	0.2%
Symantec	99%	96%	-	-	-	-
Webshrinker	98%	97%	-	-	-	-

*Retrieved through VirusTotal.

Table 4.3: Coverage for different domain popularity intervals. For each interval, we list the number of domains for which we could successfully collect labels.

The importance of being popular. Table 4.3 shows that service coverage differs depending on domain popularity.

We expect automated services to achieve a higher coverage even for less popular domains, but we observe that while McAfee and FortiGuard maintain a consistent coverage of at least 93% throughout, Bitdefender and Forcepoint drop from 93% and 98% to 27% and 48%, respectively, when labeling domains from either the top-1k or unpopular domains found in the long tail over 1M. We observe a similar behavior for Dr.Web, Websense, Trend Micro, and Alexa, who have relatively low coverage overall but perform worse for non-popular websites. The human labeling efforts of OpenDNS appear to prioritize popular domains (an expected feature). Nevertheless, OpenDNS coverage across domains ranked over the top-1M may be inflated by the 15% subdomains within that interval. As we will discuss next, in OpenDNS, subdomains typically inherit the label of the base domain. . Finally, Trend Micro (directly sourced), Symantec and Webshrinker achieve a very high coverage of over 96% for the top-10k, but their rate limits make large-scale data collection unfeasible.

In summary, only two services are able to categorize both popular and non-popular domains. Given the ever-increasing number of websites as well as the trend to conduct large-scale measurements, the choice of service impacts the capacity to classify potentially millions of visited or targeted domains, including undesired ones.

	Direct Source		VirusTotal		Intersection	
	Coverage	# Labels	Coverage	# Labels	# Labels	Consistency
Alexa	21%	1,843	33%	1,719	35	2.6%
Trend Micro	98%	75	27%	63	817	27%

Table 4.4: Differences between the results obtained through direct sources vs. VirusTotal. For this comparison, we use the top 10,000 domains in our ranked list.

Base domain vs. Subdomains. We identify 582,230 (13%) subdomains among our 4.4M domains. Three services—OpenDNS, McAfee, and FortiGuard—provide labels for more than 99% of them. Yet, as we will see in Section 4.4.3, there is no difference between base and subdomain labels in the majority of cases. In the case of OpenDNS, the improvement compared to its overall coverage (15%) stems from its approach to labeling subdomains. When humans do not offer a category for a subdomain, OpenDNS classifies it by default with the label of the base domain (if labeled). However, this coverage is skewed towards the 77% subdomains related to three base domains: `blogspot.com`, `wordpress.com`, and `tumblr.com`. For Alexa, Websense, and Trend Micro, subdomain coverage is below 1%. Depending on the source and selection of domains, overall coverage may therefore become worse.

Direct Source vs. VirusTotal. We verify labels collected through VirusTotal (which aggregates 6 existing services) by directly collecting labels for the top-10k domains at two services, Trend Micro and Alexa. As shown in Table 4.4, Trend Micro's coverage is much higher (98%) when directly queried than when using VirusTotal (28%). Moreover, only 27% of the domains are classified with the same label and only half of the distinct labels appear at both sources. As we will expand on in Section 4.4.3, we suspect VirusTotal may be using a different or an older Trend Micro product, with a potentially lower coverage and different set of labels. However, for Alexa we observe the opposite behavior: we obtain 12% more coverage through VirusTotal. Again, this may point to VirusTotal obtaining Alexa's data from an unknown source, different to our (one-time) search within the top 500 sites of Alexa's 279,716 categories. The inconsistencies between VirusTotal and a direct source indicate that the former might not be a fully reliable source. This is particularly worrisome given VirusTotal's popularity in recent academic work (Section 4.1).

4.4.3 *Labels Within Services*

In this section, we report on the distinct labels that we observe in each service, andthe properties that affect their correct and tractable interpretation: their diversity, deviations from documentation, and uniqueness.

We normalize all labels to lowercase, and we break down multi-labeled classifications into their individual units to reduce possible inconsistencies in the comparison.

Label diversity. Table 4.5 shows that the number of observed labels per service varies significantly across services, but conforms to their intended purpose. Security and content filtering services have fewer labels (12 observed in Dr.Web to 125 observed/139 documented in Forcepoint), which may simplify the setup of security policies. Conversely, the larger diversity in marketing-oriented services (300 observed/401 documented in Webshrinker, and more than 7,500 observed in Alexa) may enable more fine-grained targeting. We also see that all services except Websense use at least one label that is unique to them, showing that their taxonomies are diverse and not trivial to merge. While some services offer hierarchical taxonomies that can reduce the diversity by replacing a label with that of an ancestor, this compromises precision and forces users to decide where to prune the tree. This complexity is best exemplified by labels for Alexa queried through VirusTotal, which will only yield the label of the leaf. This is often a non-English label, derived from that website's classification into the multilingual *World* tree. For example, a given domain may be labeled as *Arts* (English), *Artes* (Spanish), or *Kultur* (German). In short, it is hard to reduce the large set of labels, without affecting their usability and interpretability.

Documented vs. Observed labels. In order to further understand how well these services document their taxonomy, we compare the documented categories with those that we observe in our dataset. As shown in Table 4.5, we observe at least one undocumented category for every service except Symantec; while Alexa doesn't explicitly document its categories, we observe only 7,557 labels for Alexa through VirusTotal, far fewer than in our own search (279,716 categories). Certain differences are due to minor syntactical variations (*e.g.,* the documented *Non-traditional religion**s*** versus the observed *Non-traditional religion_* in Forcepoint), yet they might affect researchers who search for a particular documented category and are unable to find sites within it. Other differences are due to potentially incomplete or outdated documentation. For McAfee, we still observe six categories that have been deprecated since 2010 according to their own documentation [179]. For OpenDNS, five security-related categories are unavailable for user submission or voting, as they either are restricted to trusted sources (*e.g., malware*), or appear to be legacy categories (*e.g., adware* [215]). For the Trend Micro data sourced from VirusTotal, there is a higher correspondence with its 2011 taxonomy [199] than with its 2019 one [197], suggesting that VirusTotal sources classifications from an older Trend Micro product. Finally, certain sensitive categories appear to be omitted from the documentation, *e.g., homosexuality* in FortiGuard. In summary, service documentation cannot be trusted to fully reflect the taxonomy observed in the wild, countering correct configuration and sound research usage.

Multilabeling. Six services (OpenDNS, McAfee, Dr.Web, Forcepoint, Trend Micro, and Websense) use multiple labels to categorize a single domain. This is uncommon behavior for most services, except in Dr.Web, where 67% of the domains have multiple categories, while the presence of multi-label domains is anecdotal in Forcepoint and Websense, at less than 1% of the labeled domains. Nevertheless, the number of labels that a domain can have varies for every service: in Trend Micro, 7% of multi-label domains have three or more labels,

Service	# Obs.	# Unique obs.	# Doc.	# Obs. not doc.	# Doc. not obs.
OpenDNS	64	26	58	5	0
McAfee	108	71	102	6	1
FortiGuard	87	42	86	1	0
Alexa*	7,557	7,417	–	–	–
Bitdefender*	60	34	43	25	9
Forcepoint*	125	18	139	3	21
Dr.Web*	12	6	–	–	–
Trend Micro					
through VT*					
2019 taxonomy	84	37	86	15	17
2011 taxonomy	84	37	84	7	7
direct source**					
2019 taxonomy	77	31	86	2	11
2011 taxonomy	77	31	84	9	16
Websense*	99	0	139	2	45
Symantec**	79	42	90	0	11
Webshrinker**	299	212	401	1	103

*Retrieved through VirusTotal.
**Across the top-10k domains in our ranked list. These counts are therefore lower/upper bounds of those across all 4.4M domains.

Table 4.5: Comparison of documented (*Doc.*) and observed (*Obs.*) labels, including labels unique to a particular service, across 4.4M analyzed domains unless otherwise stated.

while there is only one such domain for Forcepoint. While for other services we observe at most 6 labels for one domain, in OpenDNS, we observe `4chan.org` reaching a maximum of 17 labels.

Multiple labels may add nuance, but also complexity to their interpretation.

Next, we measure which pairs of labels frequently appear together. We observe 2,536, 1,006, 526 and 356 distinct pairs in McAfee, OpenDNS, Trend Micro and Forcepoint respectively. However, in Dr.Web and Websense, this number drops to 44 and 40; for the former, this is due to the low number of labels observed (Table 4.5). The label pairs are often unevenly distributed, *e.g.*, in Trend Micro, 2% of the labeled domains have the most popular pair *disease vector-spam*, while the next most popular pair *financial services-business economy* appears only on 0.2% of the domains. In McAfee and OpenDNS, the most popular pairs,

personal pages-internet services and *blogs-content delivery networks*, appear on 1% and 39% of labeled domains respectively. Common pairs are also not always intuitively linked. For example, in Dr.Web, the most popular pair is *adult content-social network*, appearing in 65% of all domains labeled by Dr.Web, where 60% of them are subdomains of `blogspot.com`. When using aggregated labels from VirusTotal without taking into account individual services, a non-adult blog could, therefore, be inadvertently labeled as an adult site, impacting applications targeting adult content.

Base domain vs. Subdomains. We saw in the previous section that coverage on subdomains is better compared to the general coverage, in the case of OpenDNS with an improvement of 70%. We now analyze how meaningful these labels are. We see that for OpenDNS, McAfee, and FortiGuard, 99%, 98%, and 97% of subdomains, respectively, have at least the label of the base domain. However, since domains at McAfee and OpenDNS can be multi-labeled, we observe that the percentage of the subdomains that have the same labels as the base domain drops to 46% in OpenDNS, while in McAfee, below 1% of the subdomains have different labels. This drop in OpenDNS is because 90% of `blogspot.com` subdomains, which represent 51% of the total subdomains observed, have the original label of the base domain (*Blogs*) plus an extra label, typically *Content Delivery Networks* (90% of cases).

We conclude that subdomains inherit the label of the base domain, without taking into account the actual content of the subdomain.

Labeling update. As discussed in Section 4.3, the frequency of label updates affects the timeliness and, therefore, accuracy of labels. We analyze how common such updates are for the 9 services that do not rate limit (see Section 4.3). We select 2,000 domains per service: half of them were previously labeled by (at least) that service, while the rest were unlabeled for the particular service. We select domains that have been crawled at the beginning of our data collection, to increase the time that these services had to (re-)label the domains.

We find that in our second round, only OpenDNS, FortiGuard, and McAfee categorize domains that had not been previously labeled. However, the number of updates varies: while McAfee and FortiGuard now label 88 and 53 out of 1,000 previously unlabeled domains, OpenDNS only does so for 2 domains. Similarly, for domains that had been previously labeled, McAfee and FortiGuard relabel 15 and 10 domains, respectively. The majority of these changes concern the maliciousness of domains, with some of them gaining a related label (*e.g., malicious sites*) while others lose such a label.

Finally, for OpenDNS, three domains gain a label, although two of those receive the label *Content Delivery Networks* outside of the regular voting process (Section 4.5). In summary, some services update labels over time, making it more likely that their classification better reflects the current state of a website.

4.4.4 *Labels Across Services*

The differences in both label number and coverage (see Table 4.5) call for a better understanding of the relationships between services. This analysis is how-

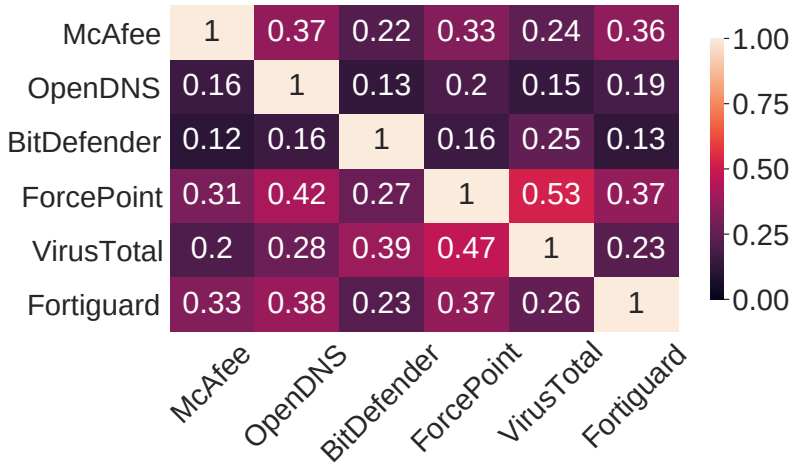


Figure 4.4: Normalized mutual information of domains with the highest degree of overlap.

ever hindered by inconsistencies in label syntax (*e.g., News* vs. *News and Media*), language (*e.g., Arts* vs. *Artes*), semantics (*e.g., File sharing* vs. *File storage*), and aggregation (*e.g., sports* vs. *entertainment/sports*). Furthermore, one provider may give multiple labels to a particular domain, requiring a comparison of sets of labels with different dimensions.

Mutual information. In this section, we take a statistical approach to perform a label-agnostic analysis. A suitable metric is the *mutual information*, which describes the amount of information gained about a random variable upon observing another random variable [216]. Mutual information can be thought of as the reduction in one variable's entropy (level of uncertainty) if the output of another variable is observed. In our case, we treat each provider as a random variable whose distribution of values (*i.e.,* labels) we estimate empirically. We can then interpret the mutual information as how similarly the labels are distributed between two services. Its normalized value will be 1 if one service assigns a common label to all domains (and none other) that are given a common label by the other, regardless of the exact label syntax. Conversely, it will be 0 if the services are completely independent, *i.e.,* there is no information to be gained about the first when observing the labels of the second.

We select McAfee, OpenDNS, Bitdefender, Forcepoint, VirusTotal and FortiGuard for this analysis as they are the services with the largest coverage (see Table 4.3). VirusTotal is a special case: while it meets the coverage criterion, its labels are aggregated from other providers, including Bitdefender and Forcepoint. The normalized mutual information matrix is shown in Figure 4.4. Overall values are low, indicating disagreement between providers, which is due to several reasons. First, services such as OpenDNS and Bitdefender differ in specialization, providing either a content- or a security-oriented label, *e.g., Online Service* vs. *Spam*. Next, human-sourced services such as OpenDNS may suffer more from subjective labeling (Section 4.5) and therefore disagree more with automated services such as McAfee. Differences in the size and granularity of taxonomies (*e.g.,* between VirusTotal and FortiGuard) can introduce further dis-

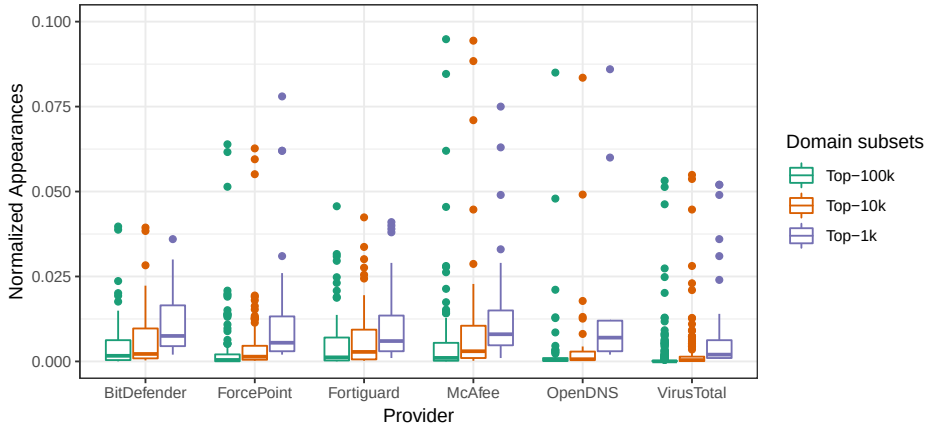


Figure 4.5: Normalized label occurrence frequencies. The statistics are computed over the number of times a label repeats itself for a given range of domains.

agreement. Finally, shared sources of labels or taxonomies may inflate agreement: we see the highest mutual information between VirusTotal and two of its aggregated providers, due to their partially shared data source. We observe consistent results when repeating our analysis using the conditional entropy.

Label frequency. Next, we compare the distribution of labels over domains, in order to understand the label coverage as well as service specialization. Figure 4.5 presents the normalized label frequencies for the top-1k, 10k and 100k domains in our ranked list. In all three subsets but in particular for the top-100k, there is a significant number of outlier labels that appear with a much higher frequency, indicating that labels are distributed unevenly. With the exception of VirusTotal, the median frequency for labels across domains is relatively consistent. On the top-1k domains, OpenDNS shows the smallest granularity in terms of coverage, while VirusTotal shows the highest. The trend is partially maintained when considering larger domain sets, where Bitdefender, FortiGuard and McAfee span the considered domains with the smallest number of labels.

Label distribution. Finally, we observe two trends in the concrete distributions of labels between providers. First, we see that, especially when considering more than two providers, one fixed set of domains corresponds to largely varying sets of labels that cannot trivially be combined into one category: *e.g.*, *Nudity*, *Society and Lifestyle*, and *Adult Content* are overlapping but not equivalent categories. We provide a visual example of these inter-service label relationships in Figure 4.6.

Secondly, we find that labels are distributed unevenly across pairs of providers: *e.g.*, for McAfee, the lower granularity of its taxonomy means that few labels cover the set of domains generated by a large number of labels from other services while for VirusTotal far more labels are needed. We look at the cumulative distribution functions of one service over a corresponding one, as we cann see in Figure 4.7. The horizontal axes contains all labels of a particular provider split into buckets, while the vertical axes represents the fraction of labels from the corresponding provider, covered by all the buckets up to the considered point. As expected, the curves for McAfee and OpenDNS (read row-wise) show

a fast increase, as a small number of buckets contains the majority of labels, while Forcepoint and VirusTotal have a much more gradual increase. In some cases, a plateau appears at a point in the curve, as in the case of the Bitdefender-Forcepoint pair, or at the very beginning, as in the case of Bitdefender-McAfee. This is an artifact of the bucketing procedure which shows that the corresponding buckets cover a very small number of labels from the paired provider. This does, however, offer interesting information regarding labels that correspond on a one-to-one or one-to-few basis, even in the case of services that have a relatively reduced amount of overall labels. In summary, differences in service purpose, taxonomy size and label distribution cause large disagreements between services, making it difficult to compare and combine their classifications.

Takeaway: *We find that commonly used domain classification services exhibit traits that affect their suitability, both for technical solutions as well as for research. Only a few services attain a level of coverage that is sufficient to cover non-popular or non-base domains. Services may return multiple or undocumented labels, requiring careful data processing and even manual validation. Breaking down multi-labeled classification may ease the label comparison between services as well as improve the interpretation of the results. However, it may also bias the results, overestimating the presence of labels that do not provide information about the real purpose of the service. The large diversity in labels, both within and across services, may harm their accurate and tractable interpretation. Efforts to combine labels from multiple services to achieve a higher agreement on label accuracy might be thwarted by labeling inconsistencies. The labeling updates may also have an impact on accuracy and timeliness. Researchers should be aware of these phenomena and renew their dataset to reduce possible misclassifications, especially in treating malicious services. In summary, sound deployment and usage of domain classification services requires a thorough understanding of the (desired) characteristics and resulting biases to select the most appropriate sources.*

4.5 HUMAN PERCEPTIONS

As described in Section 4.3, OpenDNS, DMOZ and Curlie leverage a network of human volunteers to label domains. In OpenDNS, moderators approve or reject labels voted on by users, while in DMOZ and Curlie, editors add suggested sites to their managed categories. In this section, we harvest historical data from OpenDNS' voting process to further measure the effect that human decisions have on (1) OpenDNS' labeling process—in terms of user and editor temporal dynamics—and, (2) on the resulting classifications. For comparison and completeness, we also study Curlie's labeling dynamics by crawling and analyzing their publicly available data.

4.5.1 *Labeling Dynamics*

OpenDNS. OpenDNS relies on a voting process that allows users to submit labels ('tags') for domains, which then receive positive and negative votes from other users. After sufficient votes, a trusted moderator approves or rejects these

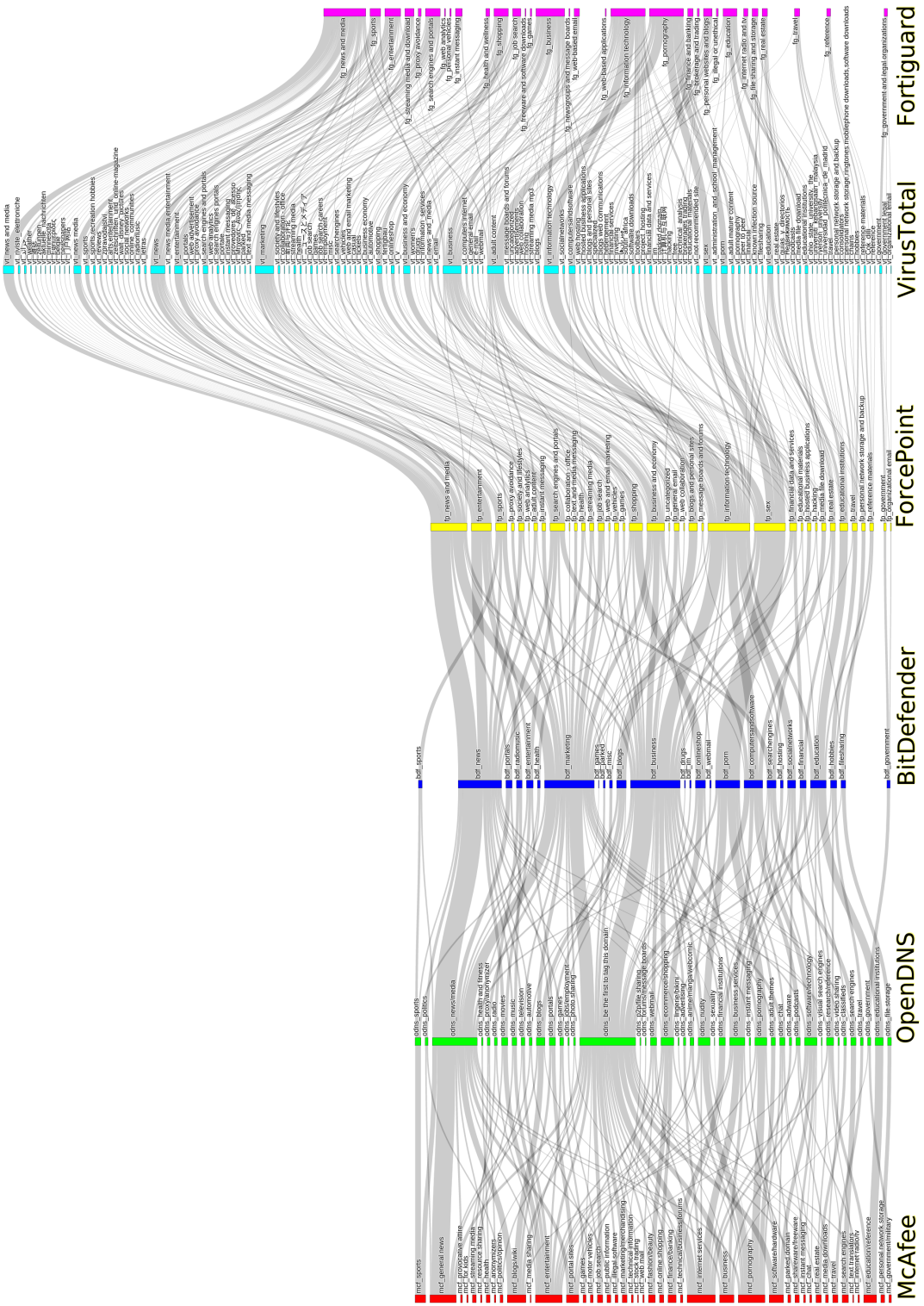


Figure 4.6: Label correspondences from top-1k domains for McAfee, OpenDNS, Bitdefender, Forcepoint, VirusTotal and FortiGuard.

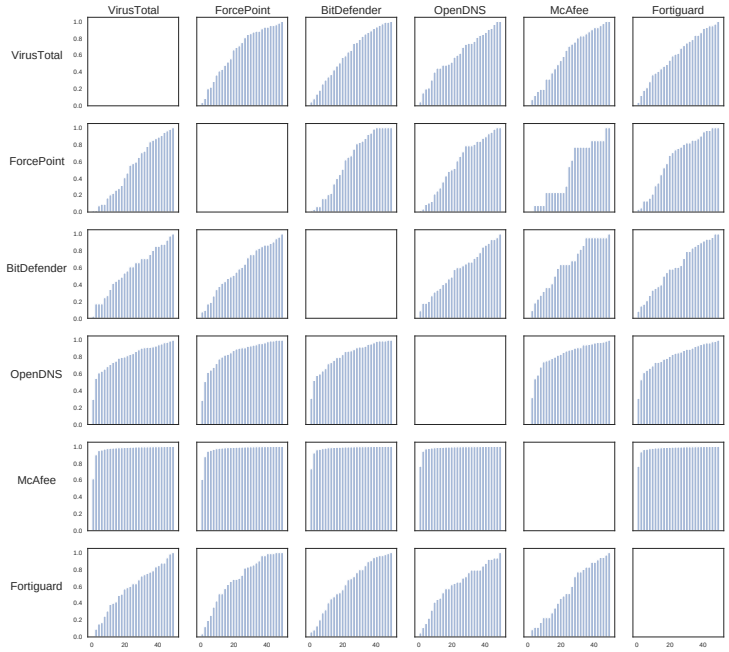


Figure 4.7: The distributions of labels for the six providers show considerable variation. Each row of the matrix represents the coverage of one provider in terms of the corresponding provider on the column. McAfee, Bitdefender and FortiGuard have a relatively small number of labels covering the set of domains, compared to the finer granularity of VirusTotal or Forcepoint. As to one label of McAfee, for example, there corresponds a considerable number of labels from VirusTotal, the conditional probability between pairs of labels from the two services is small, explaining the low values of conditional entropy as well as low mutual information. This is valid in all such one-to-many correspondences between providers.

submitted labels [176]. OpenDNS publicly releases historical data from this voting process, including the labels proposed for every domain, the user who proposed them, whether they are accepted or not, and the moderator who took the final decision. All items are timestamped, which allows us to analyze the evolution of submitted labels over time. This data allows us to inspect the OpenDNS voting process for 794.8k domains, as well as the behavior, agreements and disagreements between 19k users and 292 moderators from February 2008 until January 2020.

First, we analyze who is submitting labels for observed domains. The first observation that stands out is that most users are "casual," as 95% of users only submit a label for 10 domains or fewer. Nevertheless, there is a group of 160 highly engaged users who submitted labels for more than 100 domains. As for moderation, the workload distribution is more even: around 40% of moderators have approved 10 labels or fewer. Nevertheless, there are 292 moderators (0.03%

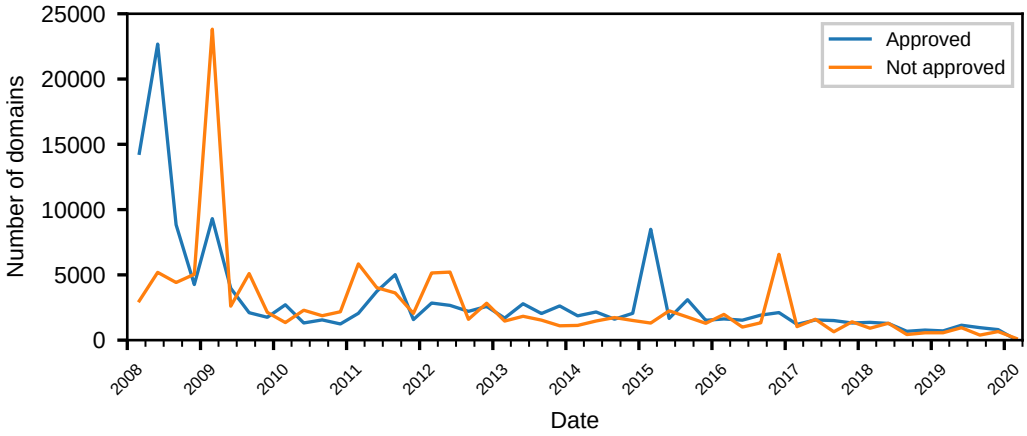


Figure 4.8: Domains labeled in OpenDNS by quarter.

of all moderators) which are very prolific, being responsible for the approval of over 10k labels.

Figure 4.8 shows the number of approved and not approved labels submitted quarterly. We can observe that the majority of labels were submitted during 2008 and 2009. Interestingly, at the beginning, the majority of labels were accepted. However, starting in 2009 there is a large decay on the number of accepted labels and an increment of those that are not accepted. Our intuition is that because at the beginning of the project all major sites lacked a label, the probability of people correctly labeling those is higher. As time passes, only a long tail of unpopular domains remain unlabeled, so users are more likely to submit an incorrect label or no label at all.

Curlie. As in the case of DMOZ, Curlie has no open voting process. Instead, trusted editors fully manage categories and decide which user suggestions they include. Review may come from other editors for the same category and its parent categories, or those with the right to edit all categories [209, 217]. Because of its content discovery purpose, Curlie has a large and deep hierarchical taxonomy, consisting of 671,715 observed categories. By analyzing the assignments of categories to editors, we examine whether these editing and reviewing processes can be effective considering this deep taxonomy.

Only 985 (0.1%) are explicitly managed by at least one out of 294 active editors. When we account for the editing rights to subcategories, 515,791 (76.8%) categories have *at least* one "implicit" editor.

However, 565,812 categories have *at most* one implicit editor, which means that 84.2% of categories can only be peer reviewed by the editors with rights to all categories. The opportunity for peer review may be further affected by the breadth of certain editors' scope, with the top "implicit" editor managing over 300k categories. In summary, the large number of categories managed by only a few editors may prevent these editors from conducting a regular review for accuracy and recency.

Figure 4.9 shows that around half of all categories have been updated since the evolution of DMOZ into Curlie in 2017. Moreover, it shows more recent activity higher in the tree: lower levels may either inherently require fewer up-

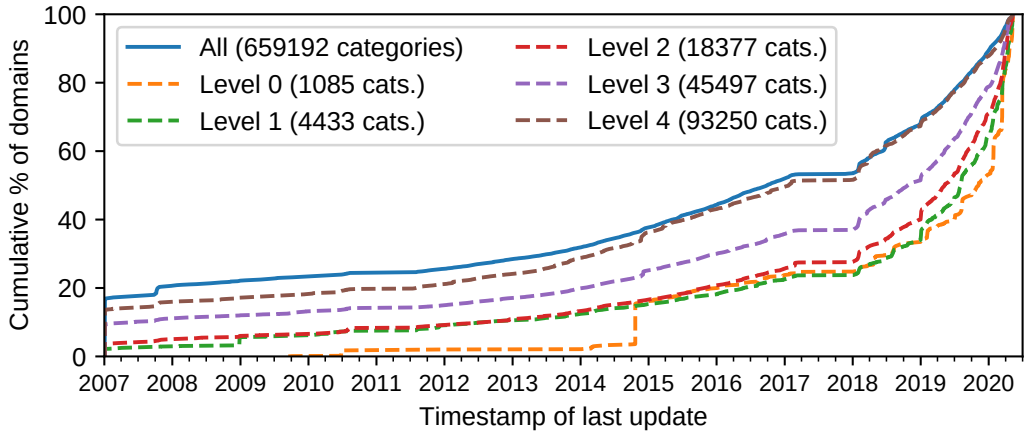


Figure 4.9: Cumulative distribution of update timestamps for categories in Curlie.

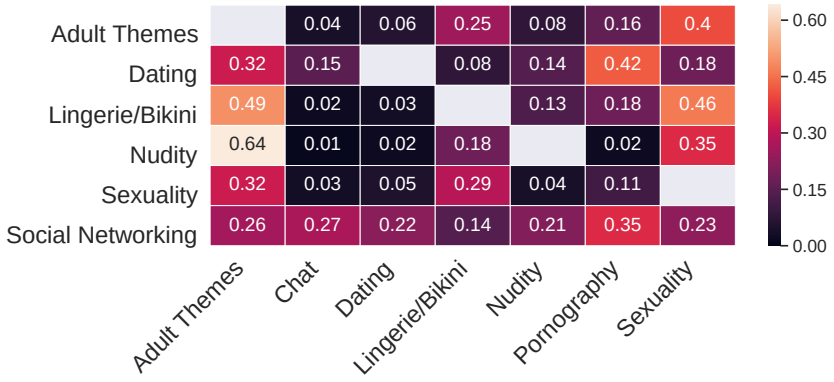
dates, or may be less actively maintained by their editors. While there is steady ongoing activity on Curlie, many categories have not been updated for years, potentially leading to their entries being outdated or inaccurate.

4.5.2 *Labeling (dis-)agreements*

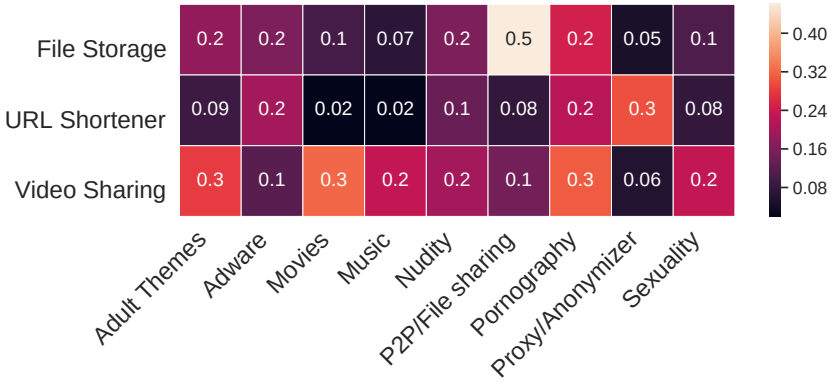
One key issue with human-in-the-loop labeling is that the task of classifying domains is not completely objective, and thus different users might suggest different labels for the same website. Therefore, we measure how often this happens in the labeling process of OpenDNS. While the median number of accepted and rejected labels in OpenDNS is one, we have shown in Section 4.4.3 that some domains have as many as 17 accepted labels. In the case of labels that do not get approved, we can find domains with a high level of disagreement among voters with as many as 58 not accepted labels.

We further investigate the type of labels that create most agreement and disagreement in OpenDNS. To do so, for all domains with a given approved label, we measure how often other proposed labels are approved and rejected for the same domain. Selected clusters of labels where the disagreement is high are shown in Figure 4.10. Some of the labels that often appear together seem to be a product of honest mistakes by the users, as they are closely related (such as *Adult themes* and *Sexuality*, or *Travel* and *Business Services*).

An interesting case is the label *Pornography*, which often appears proposed (and rejected) in addition to other labels. While this might make sense for some categories (such as *Lingerie* or *Sexuality*), it is surprising that over 30% of *Social Media* sites and over 40% of dating sites were also labeled (and rejected) as *Pornography*. Another apparent issue is that domains related to *URL Shorteners*, *Video Sharing* or *File Storage* can often be related to other categories, such as *Music*, *Movies* or *Pornography*. This shows that deciding the correct label for a given domain can be hard, with the differences between categories being vague. Furthermore, not all users might behave honestly, as some could mislabel domains to pollute the system or gain advantages over competitors, *e.g.*, a pornographic



(a) Adult content cluster



(b) Internet utilities cluster

Figure 4.10: Examples of overlap between categories in OpenDNS. The heatmap shows the frequency of X-axis categories being rejected when the Y-axis category is approved.

site trying to be labeled as a video sharing site, or a company labeling a competitor's website as malicious or pornographic.

In the case of labels that often appear accepted together, we also find a high correlation among categories that could be related to sexual or nudity content (*e.g., Pornography*, *Nudity*, *Bikini/Lingerie*). Another interesting case is the pair *Advertising* and *Business Services*, which are accepted together over 30% of the time. This can be a result of many of these *Business services* acting as third parties offering advertising and tracking services too. Similarly, *News and Media* and *Television* often appear together since television stations often act as news outlets.

4.5.3 *Is labeling domains a trivial process?*

We perform an experiment using the authors of this study in order to gain a better idea of the aforementioned challenges behind OpenDNS' labeling process. One member of the research team manually selected 200 hostnames, including 50 for which OpenDNS and McAfee provide semantically equivalent labels; 50 for which they disagree; 50 from the top-1k domains in our normalized rank; and 50 unpopular sites. For ethical reasons, we discarded domains

with labels that could be uncomfortable or harmful for our human labelers (*e.g.*, child pornography, nudity, violence, drugs, weapons, and malware-related ones). The remaining authors manually visited each website and labeled it using the OpenDNS taxonomy and definitions. Each domain was labeled by two authors, adding a third labeler when there was disagreement in the first stage.

Disagreement between two labelers is relatively high at 35.5% of domains, reaching 90.5% agreement between at least two reviewers when a third labeler is introduced. When the final results are compared to OpenDNS categories, we observe that our process could only achieve 71% accuracy; in 80.5% of the cases, at least one labeler reported the same category as OpenDNS. This experiment, while not representative, illustrates some of the challenges that arise when humans are involved in the process, even for experts in network measurements and cybersecurity. Disagreement is the result of subjective factors caused by different perceptions and sensitivities, but also by the inherent ambiguity of many of the categories forming the taxonomy and the dual nature of many websites, for instance, blogs offering political content [218] or tourism boards advertising casinos [219].

Takeaway: *We analyze OpenDNS's ecosystem of voters and editors, and find that most labels were submitted during the early stages of the project. We show that most users (95%) submit labels for only a few domains but that, in general, workload is evenly distributed among moderators. In the case of Curlie, we find that peer review may suffer from the low number of editors, but that categories are still being updated regularly. Furthermore, we find that labeling strategies involving humans are bound to generate disagreements. In OpenDNS, there are domains with 58 not approved labels. Moreover, the slight differences among labels generate clusters of related labels that often appear rejected together (i.e., Adult themes, Lingerie/Bikini, Pornography and Sexuality). We show that labeling is a non-trivial job by running a small-scale manual classification experiment, in which we only achieve 71% accuracy compared to OpenDNS and find that two labelers disagree on 35.5% of domains—highlighting the subjective nature of labeling.*

4.6 CASE STUDIES

In this paper, we have shown that researchers often rely on domain classification providers to either understand the type of domains that they observe in their study [155] (*i.e.*, to better characterize their results) or to gather a field-specific corpus of domains, including the corpus we use on Chapter 3.

Next to that, core applications of domain classification services are outside the academic circles. They are often used in technical solutions for content filtering and threat intelligence, for example in parental control apps [220] and school networks [140], which require accurate identification of specific types of domains.

Therefore, in this section we aim to understand whether choosing one domain classification service over another can yield different results when selecting target domains or when classifying domains specific to a given category. We analyze the usefulness and aptness of domain classification services for three types

Type	Ad/Tracking N=24,825		Adult content N=3,519		CDN N=2,858	
Label	Any	Related	Any	Related	Any	Related
OpenDNS	16.9%	3.9%	88.2%	88.0%	3.4%	0.2%
McAfee	70.8%	3.7%	99.1%	97.6%	98.2%	84.7%
Fortiguard	78.7%	7.7%	99.8%	98.8%	93.0%	81.7%
Alexa	2.8%	0.1%	1.1%	0.1%	0.1%	0.1%
BitDefender	32.0%	2.8%	83.5%	65.8%	27.0%	27.0%
Forcepoint	51.6%	15.1%	97.1%	94.9%	29.9%	3.1%
Dr.Web	9.0%	0.0%	92.5%	92.4%	0.3%	0.0%
Trend Micro	7.4%	0.9%	12.1%	11.8%	0.4%	0.0%
Websense	2.8%	1.0%	4.6%	4.5%	0.2%	0.1%

Table 4.6: Coverage for different types of domains.

of domains that are often analyzed by the research community: (1) advertising and tracking services; (2) websites offering adult content (*i.e.*, pornography and gambling sites); and (3) domains that belong to a Content Delivery Network (CDN) and hosting providers. Our approach starts with obtaining available sanitized domain category sets to identify which domains belong to each one of these categories. Then, we analyze the coverage as well as the labels assigned to these domains by different classification services to identify potential errors and inconsistencies. While such specialized lists are more appropiate for choosing a pool of websites that belong to a given category, we have seen that it is still common for academic papers to rely on classification services for website selection or classification [168, 156, 149].

Advertising and tracking services. As ground truth, we take a list of manually sanitized domains indexed in EasyList [84] and EasyPrivacy [221].[5] However, these lists allow blocking traffic at a full URL level.[6] To reduce bias in our case study, we opt to account only for domains that are fully blocked by these lists, regardless of the full URL path.

After a manual sanitization process, we study the labels from different classification services for the resulting 24,825 advertisement and tracking-related domains and manually extract the resulting labels semantically related to advertising and tracking applications (*e.g., Web Marketing* or *Advertisement*).

Table 4.6 (two leftmost columns) shows that none of these services are able to correctly label most domains as tracking or advertising. Forcepoint presents the highest accuracy, which is barely higher than 15%, at the cost of sacrificing

5 Both used by the anti-tracking solutions AdBlock and AdBlock Plus [101, 102].

6 *e.g.*, they would not block the `bbc.co.uk` webpage, but they would block any URL from this domain which contains the `tracker.js` file [222].

coverage (51.6%). While McAfee and FortiGuard have a higher coverage, they classify fewer than 10% of the domains as trackers.

Most of the errors arise from tracking- or advertising-specific subdomains. For instance, all providers classify `airpushmarketing.s3.amazonaws.com` and `tracking.eurosports.com` using labels related to *hosting/CDNs* and *news/media/sports*, respectively.

Identifying adult content. We rely on two resources to gather domains related to adult content [140]. First, we rely on a manually labeled and sanitized list of pornographic websites we use on the Chapter 3. Additionally, we compose a list of gambling sites extracted from three government websites [223, 224, 225]. By combining these two sources, we compile a manually vetted list of 3,519 domains related to web services typically considered as "adult content".

The results (Table 4.6, middle columns) show that 5 services do a good job at identifying and correctly labeling webpages that host adult content: OpenDNS, McAfee, FortiGuard, Forcepoint and Dr.Web. Yet, there are substantial differences across services. Alexa, Trend Micro and Websense do not provide a label for the majority of the websites analyzed. Therefore, this case study also demonstrates that the choice of one provider above another can have severe implications in the number of domains classified as adult content. We also examine which other labels are usually assigned to adult content domains, finding a high correlation with those related to video sharing and streaming media. These labels are, in most cases, technically correct but they do not allow to identify these domains as pornographic. We also see that some services assign labels that imply maliciousness of adult domains (*e.g., malicious*, *spam*, or *not recommended*).

CDN and hosting provider related domains. Content delivery networks (CDNs) remain the dominant means for serving popular content and represent Internet *infrastructure*. While most domain classification services (*e.g.,* McAfee and Fortiguard) contain labels referring to CDNs or hosting providers, the *content* classification is often mixed with an *infrastructure* classification. As an example, one service can classify a CDN-hosted site as *content delivery network* while another derives a label from the site's content (*e.g., news* or *personal blog*).

In order to measure differences in the classification strategies of different services, we select those domains in our dataset that are related to CDNs and hosting services. To do so, we pattern match the CNAME record of all domains against more than 80 CDN signatures from WebPageTest [226]. In total, we obtain a corpus of 2,858 domains, for which we compare the coverage across domain classification services. Table 4.6 (rightmost columns) shows that only McAfee and FortiGuard provide a label for the majority of these domains. Both services classify these domains based on their function rather that on their content (*e.g., Internet Services*, *information technology*, and *content services*).

For the other services, the coverage is so low that it is difficult to discover a trend in the labels. Yet, it is still possible to find examples of labels related to the actual content of webpages hosted on these services (*e.g., News*, *Adult content*, or *Business*) as well as to the type of service provided. None of these classification strategies are right or wrong, but the choice of service translates in differences in terms of coverage and labels for CDN and hosting provider related domains.

Takeaway: *For specialized use cases, the choice of one domain classification service over another can significantly impact the accuracy of academic studies and the effectiveness of solutions relying on them.*

4.7 DISCUSSION

In this section, we extract actionable insights from our empirical results, discuss best practices for using domain classification services, and propose various solutions as future work to overcome their limitations.

Dealing with insufficient accuracy. The key observations of our study are that *i)* coverage varies substantially between services (Section 4.4.2) and *ii)* the classification accuracy is marred by inconsistent taxonomies (Section 4.4.3) and low agreement among providers (Section 4.4.4). These inherent limitations set a high barrier for their effectiveness in real-world applications as well as their usage in research.

For highly targeted use cases, general-purpose classification services may fall short. For example, as shown in our case studies (Section 4.6), the choice of service impacts the number of correctly identified adult domains. It may therefore be necessary to either search or develop curated and manually labeled domain-specific lists. Furthermore, end users and researchers should carefully consider the implications of errors. In applications like content filtering, errors can lead to inappropriately restricting access to legitimate resources ('overblocking') or, conversely, allowing access to undesirable resources ('underblocking') [227, 228]. For example, aggressive adult content filters could block sexual health information [229] or, as in the recent case of Cloudflare's DNS resolver, LGBTQIA+ sites [230]. In the academic domain, researchers can also take into account how important classification is to their studies, *e.g.*, using domain categories to provide context for a minor result vs. generating the list of domains on which they base their whole study. There are a few documented cases in which authors preferred their own classification over those of commercial services due to concerns regarding their accuracy and coverage [37, 169, 36, 95, 153].

Dealing with biases. Coverage and accuracy suffer from selection and interpretive biases respectively. Service purpose determines which and how domains are classified: a filtering service may better cover and differentiate malicious domains, while marketing- or discovery-oriented services may provide a more fine-grained label for popular sites. How labels are sourced also introduces biases. For automated solutions, these stem from deficiencies in the training sets for machine learning algorithms. In a manual classification process, these are induced by maintainability challenges as well as human interpretation (Section 4.5). There are cases where using a domain classification service can produce sound results. Yet, researchers should gain a proper understanding of potential biases in their chosen services to assess the limitations of applying them in specific domains, *e.g.*, by consulting the documentation. To empirically gauge the coverage and accuracy of the used service specifically for their studied domains, researchers can additionally manually inspect random subsets to

determine whether the labeling is of sufficient quality to make its usage appropriate.

Dealing with inconsistencies. When using domain classification services, results must be interpreted and reported with care, to avoid introducing errors due to inconsistencies. Domain classification services exhibit varying characteristics, *e.g.*, whether they provide multiple labels, label subdomains differently, or regularly update labels (Section 4.3). Moreover, they may behave unexpectedly, such as by deviating from their documented taxonomies (Section 4.4.3). Users should therefore verify the output of the services, *e.g.*, by analyzing aggregate statistics or a randomly selected sample. Furthermore, the specific applications of services affect their taxonomies. The granularity and exact meaning of a label (even if it is syntactically the same) thus largely differs between services and directly impacts the effectiveness of any application or the results of any study. Studies based on domain classification should thus examine the labeling taxonomy in detail and report the meaning of the selected labels to prevent wrong or incomplete conclusions.

Aggregation of multiple domain classifiers. Many websites are complex entities: it is hard to reduce them into a single label. Researchers might be tempted to overcome the limitations of individual domain classifiers—both in terms of coverage as well as label accuracy—by combining the output of multiple services in a single analysis pipeline. While this might be useful in some scenarios (*e.g.*, threat intelligence aggregators such as VirusTotal), we identify multiple challenges that rule out simplistic aggregation strategies:

(1) If the goal is to improve overall coverage, aggregating various classifiers might not necessarily achieve this purpose, as we showed in Section 4.4.2. The choice of classifiers should be informed by the size of the intersecting set. In addition, we found coverage to vary greatly depending on factors such as domain popularity or freshness.

(2) Different classifiers might provide complementary perspectives on a domain's nature, but the aggregation of their labels can be difficult since they come from different taxonomies with radically different purposes. Simply taking the union of the outputs might unnecessarily increase the constellation of labels and increase redundancy, since two services might use semantically-equivalent labels to reflect the same purpose or abstract concept. This could be aggravated by services developing multilingual taxonomies. Reconciling multiple taxonomies coherently might be cumbersome and difficult to scale, particularly if it must be done semantically.

(3) Determining what is a discrepancy among classifiers and what is just a different perspective on the nature of a website could also be challenging. A site can simultaneously be labeled as *porn*, *streaming*, and *CDN* by three different providers. Understanding the focus, sensitivities, limitations, classification methods, and intended label usage of each classification service is an unavoidable step to properly contextualize and meaningfully aggregate their outputs.

Part IV

AUDITING PROFILING AND AD TARGETING ALGORITHMS.

5

AUDITING PROFILING AND AD TARGETING ALGORITHMS.

The main advantage of online advertising compared to other mainstream advertising channels (TV, radio or newspapers) is its capacity to deliver personalized ads. The ad tech industry has developed a sophisticated tracking ecosystem to create accurate profiles of each user, including demographic characteristics (*e.g.,* gender and age), location information (*e.g.,* home town or current location), and interests (*e.g.,* cars, sports, food & beverage, *etc.*). Then, advertisers can configure advertising campaigns targeting users with particular characteristics, Targeted ads are meant to reach users whose inferred profiles meet the definition of the targeted audience.

Therefore, the performance of advertising campaigns depends on the capacity of the online advertising ecosystem to properly infer users' profiles. Inaccurate profiling algorithms may lead to severe damage to advertisers: 1) wasting advertising budget on users that are unlikely interested in their products or services; 2) annoying users with irrelevant ads, which may contribute to increasing the use of ad blocking software [231, 232]. A second important aspect related to users' profiling is its implications in the context of citizens' data protection and privacy. This has led to the development of new regulatory frameworks in the area of data protection, like the European General Data Protection Regulation (GDPR) [1], the future ePrivacy regulation [30] and the California's Consumer Privacy Act (CCPA) [2].

Despite the obvious importance of the accuracy of profiling algorithms for advertisers and users alike, the conventional wisdom from researchers, practitioners, and public institutions, which focus on the presence of tracking mechanisms [95, 24, 233, 13], and the regulatory compliance of such services, seems to assume that profiling algorithms used in online advertising perform well, and thus the created profiles are fairly accurate. Indeed, there is just a previous work by Bashir *et al.* [172] addressing this problem.

In this work, we propose a pioneering analysis on the accuracy of profiling algorithms in online advertising and their impact on the performance of ad targeting algorithms. We focus our work on analyzing Google and Facebook (FB), the two most important companies in the online advertising sector, accounting together with ~54% of its market share [234].

In particular, we aim to answering the following questions:

- How accurate are the users' profiles constructed by Facebook and Google?
- How accurate are the targeted ads delivered by Facebook and Google's ad campaigns?
- What is the answer to the previous questions for the case of sensitive and socio-demographic data?

To this end, we build a measurement methodology that leverages a chromium browser add-on extension, available for Chrome, Brave, and Edge. Our browser extension collects the user's profile information for Facebook and Google periodically from the transparency tools that both platforms offer [235, 236]. In addition, the extension allows users to fill a survey in which they can rank each of the attributes obtained from their profile on a scoring scale from 1 (very inaccurate) to 5 (very accurate). Finally, the browser add-on extension collects all the ads shown to the user by Google and Facebook in the browser instance where the add-on is installed, as well as the explanations offered by Google and Facebook of why the user has received each ad [237, 238]. Our browser extension has been installed by 62 users, who have provided 6,400 responses about their Facebook and Google profiles. Moreover, we have collected 193,842 ads delivered by Google and Facebook to these users.

By processing the collected information, we can address the questions presented above. To this end, we compute two metrics: 1) *Profiling Accuracy* defined as the distribution of scores (between 1 and 5) given by the users to their profile attributes, which allows analyzing the accuracy of users' profiles; 2) *Targeting Accuracy* defined as the distribution of scores (between 1 and 5) assigned by users to the targeted attributes associated to the received ads. This metric serves to assess the performance of ad targeting algorithms from Facebook and Google. Finally, 3) we compute these two metrics for those attributes that we manually classify as socio-demographic or potentially sensitive, to study the accuracy aspects related to profiles' socio-demographic and sensitive attributes. The analysis of our dataset reveals the following main findings:

1. Google and Facebook offer a poor *Profiling Accuracy*. In particular, 50% and 47% of the user attribute on Facebook and Google profiles, respectively, are not accurate, according to the users' scores (*i.e.,* these interests are assigned a score of 1 or 2). On the other hand, just approximately 1/3 of the attributes in both Facebook and Google receive a score of 4 or 5 and thus can be considered accurate. This result raises serious concerns about the accuracy of profiling algorithms.

2. Google shows a similar distribution for the *Profiling Accuracy* and *Targeting Accuracy*, indicating a poor performance of both profiling and ad targeting algorithms. Instead, Facebook presents a significantly better *Targeting Accuracy* (over 51% of targeted ads include categories with scores 4 or 5) compared to Google.

3. We observed that 55% and 87% of Facebook and Google users in our dataset present sensitive attributes in their profiles, respectively. In particular, our data indicates that Facebook assigns, on average, 11 sensitive attributes to users, compared to just 2 in the case of Google. Further, we saw that Facebook assigns incorrectly sensitive interests, particularly 42% of the responses on Facebook received a score of 1 (very inaccurate), while only 5% received a score of 5 (very accurate). This suggests that sensitive attributes are generally wrongly inferred by Facebook profiling algorithms.

5.1 BACKGROUND

In this section, we provide an overview of fundamental concepts needed to understand this chapter. We provide an overview of the tracking & profiling techniques, the ad delivery methods, and the ad transparency tools used by Google and Facebook.

5.1.1 *User's Profiling*

Online advertising services have developed a sophisticated tracking ecosystem that allows them to record the online activity of users. The information captured from users gives them the capacity to create users' profiles and distribute personalized ads. In this section, we describe how Google and Facebook create the profiles and highlighting the main differences between both platforms.

Google: The basic demographic information of Google users is typically self-reported by them when creating an account on Google. Indeed, during the creation of a Gmail account users are requested to provide their name and surname, birth date, and, optionally, their gender and their cell phone number. Additionally, Google tracks the activity of users in its proprietary platforms, operating systems, and services such as Google Search, Gmail, Google Maps, Youtube, Google Chrome, or the Android OS for mobile devices [239]. In addition to the tracking conducted in these venues, research works have revealed the presence of Google as a third-party in a large fraction of Internet websites and mobile Android apps [95, 24]. Finally, Google also tracks the activity of users in the real world by collecting users' mobility patterns, and the locations visited, through the Android mobile OS or applications such as Google Maps. All this rich information serves as input data to Google's profiling algorithms, which, based on it, infer the interests of users. In particular, Google uses a hierarchical taxonomy [240] that includes the categories assigned to users' profiles. In addition to the self-reported data and the categories explicitly declared in its taxonomy, Google can infer additional socio-demographic and sensitive information about users. Indeed, a legal complaint has been filed in Europe for possible massive leakage of sensitive data associated with Google's audience taxonomies [241].

Facebook: The main sources of data for Facebook are its social media platforms, Facebook, Messenger, and Instagram among others. As in the case of Google, the basic demographic information available for profiling algorithms is self-declared by the user. During the registration process on Facebook, the user is required to report name and surname, mobile phone or email address, and birth date, while gender is an optional value (in the case of Instagram, less information is required). Additionally, Facebook tracks the activity of users in these platforms, *e.g.,* posts they create, posts they like or click, comments, followed groups or accounts, etc. Through the use of sophisticated Natural Language Processing algorithms and tagging systems, they can map the actions of users into indications of preferences or interests on certain topics [242]. Similar to Google, Facebook is also present in an important fraction (although smaller

than Google) of websites and mobile apps, using third-party cookies installed on browsers and SDKs on mobile apps [95, 24]. Moreover, Facebook's mobile app serves to track the location information of users to infer mobility patterns and visited places. Like Google, Facebook uses a predefined taxonomy of categories to assign attributes to users' profiles and it also infers socio-demographic (*e.g.,* the level of education) and sensitive (*e.g.,* the sexual orientation [46]) information about users. The main difference between both platforms is that the Facebook taxonomy is very granular and includes millions of categories. A subset of the main categories of the Facebook taxonomy can be found in [243].

5.1.2 *Ads delivery*

Advertising platforms allow advertisers to configure targeted ad campaigns, which target users based on a pre-defined set of attributes. In this section, we explain the ad delivery processes considered in this paper for Google and Facebook.

Google: It has a predominant position in the online advertising ecosystem, which allows it to distribute ads through different channels. These span from its platforms (like Youtube), in which Google owns and manages all the ad spaces, to ad spaces in webpages and mobile apps owned by third parties, typically referred to as *publishers*. In the context of this paper, we analyze the operation of Google to serve ads in the latter case. Publishers owning a web page or a mobile app offer ad spaces, which typically are handled by a third party. Indeed, Google is the company responsible for handling a major portion of publisher-owned ad spaces in websites and mobile apps. Let us explain the ad delivery process with a simple example for the case of a webpage (note that the case of mobile apps is similar). When a user visits a webpage, the ad space (typically embedded in an iFrame) sends a request to the entity handling it, Google in our case. Google compiles all the possible information about the ad space: 1) information of the space itself (size, allowed type of ad, position in the page, etc.); 2) information about the browser, operating system, and type of device (mobile vs. fixed); 3) information about the user visiting the website, *i.e.,* the user's profile. At this point, Google looks for an advertising campaign whose audience matches the user's profile, from among the advertisers configuring their campaigns on Google's advertising platform. Typically, there is more than one campaign matching the offered profile. Then, Google runs an auction process to choose the ad campaign whose ad will be delivered.

Facebook: The Facebook advertising ecosystem operates as a walled-garden, in which ads are delivered via the social media platforms owned by Facebook: like Facebook itself or Instagram. Note that a minor part of Facebook's advertising business is dedicated to delivering ads in third-party venues (mobile apps or webpages). In the context of this paper, we only focus on the ads delivered in the Facebook social network. We discard Instagram since it is mainly used through the mobile app. Furthermore, after a manual validation, we observed that the browser version of Instagram hardly offers any ads. When users open the Facebook app in their mobile phone or web browser, they find two types of

ads: i) ads that appear as a post on the main wall, identified as "Sponsored", and ii) ads on the right upper corner of the screen, under the "Sponsored" tag (only available in the web version). Each of these spaces is treated as an ad space. Similar to the case of Google, when a user opens their Facebook account and ad space is loaded, Facebook collects the profile information of the user and seeks an ad campaign that matches the user's profile. Usually, several campaigns would be targeting the user's profile, and thus an auction process is run by the Facebook advertising platform to select the ad campaign whose ad will be delivered into the ad space.

5.1.3 *Transparency Tools*

The controversial cases about the use of privacy-invasive practices used by online advertising platforms have put the focus on big tech firms, especially on Facebook and Google. As a consequence, these tech firms have reacted by offering *Transparency Tools*. There exist two types of such tools: *Ad Preference Managers*, which allow users to control and verify the data assigned to their profiles, and *Ad Transparency Tools* which inform users about the targeting parameters associated with the ads delivered to them.

5.1.3.1 *Ad Preference Managers*

An ad preference manager allows users to access their profiles in an advertising platform and modify them at their will, removing interests and sociodemographic attributes.

The ad preference managers of Google and Facebook are referred to as *AdSettings* [235] and *Facebook Ad Preferences* [236], respectively.

5.1.3.2 *Ads Transparency Tools*

These services provide users with an an explanation of why the user has received a specific ad. The Ads Transparency tools of Facebook and Google are referred to as *Why Am I seeing this Ad?* [238] and *About Ad* [237], respectively.

Facebook's *Why Am I seeing this Ad?*: This tool provides a very detailed explanation associated with each ad received by a user. In particular, it offers the complete list of attributes of the targeted ad campaign associated with the ad. By cross-checking the attributes of the user's profile and the attributes of the targeted ad, users can easily infer the specific attributes for which they have been targeted.

It is important to highlight that Facebook uses the same set of categories for the users' profiles and the targeting options an advertiser can configure in its ad campaigns.

Google's *About Ad*: This tool provides significantly less detailed information about the reasons why a user has received an ad compared to Facebook's tool. Google uses a set of 26 pre-defined high-level reasons in the *About Ad* tool. Table 5.1 shows them. The explanation offered by Google for a targeted ad is formed by a combination of one or more of these 26 pre-defined reasons.

Explanation	Ads	User's Profile
Your visit to the advertiser's website or app	10923	
Websites you've visited	5652	
The time of day or your general location (like your country or city)	3743	
Google's estimation of your interests	2969	✓
Products often shown together	2328	
The information on the website you were viewing	2067	
Information in your Google Account	1527	✓
Google's estimation of your age group, according to your activity while you were signed in to Google	1184	✓
Information collected by the publisher. The publisher partners with Google to show ads	999	✓
General factors about the placement of the ad, agreed upon by the publisher (ex: website, app) and the advertiser	999	
Your similarity to groups of people the advertiser is trying to reach, according to your activity on this device	955	✓
Your age group	766	✓
Your similarity to groups of people the advertiser is trying to reach, according to your activity while you were signed in to Google	709	✓
Google's estimation of your gender, according to your activity while you were signed in to Google	312	✓
Popular products from this advertiser	260	
Google's estimation of your Parental Status	249	✓
Your gender	238	✓
Google's estimation of your interests, based on your activity while you were signed in to Google (for example, your searches)	61	✓
Google's estimation of your Education Status	15	✓
The advertiser's interest in reaching new customers who haven't bought something from them before	13	✓
Information you gave to the advertiser, which the advertiser provided to Google. Learn more	12	✓
The website you're on	6	
The time of day	6	
Your general location (like your country or city)	6	
Google's estimation of your Marital Status	5	✓
Google's estimation of your Homeownership Status	3	✓

Table 5.1: Google's About Ad explanations, ranked by the number of ads in which each of them appears.

We have manually identified which of them (16) may be related to information available in the end user's profile and marked them in Table 5.1 (column "User's Profile"). Overall, these reasons are very generic, and thus they do not provide a sufficient level of detail to know the actual targeting attributes used by the advertiser.

5.2 METHODOLOGY

We implemented our methodology to analyze the accuracy of online profiles and the performance of the ad targeting algorithms. This methodology consists of several steps that we describe in the following sections:

5.2.1 *Add-on implementation*

The first challenge to conduct this study is to obtain data from real users, including the attributes associated with profiles and the ads received. To address this challenge, we have implemented a browser add-on extension that operates in any chromium-based browser, including Google Chrome, Microsoft Edge, and Brave. The extension collects the following information:

Collection of end users' profile information: One of the main goals of our add-on is to collect the attributes of users' profiles from their Google and Facebook accounts. When the user installs the add-on, it detects whether the user is logged into Facebook and Google. If the user is not logged, the add-on shows a pop-up asking the user to do so.

Then, the add-on connects to Google's Ad Setting webpage [235] and Facebook Ad Preference webpage [236] (in the background) from where it collects the profiles attributes scrapping the HTML. The add-on repeats the process of collecting the attributes from profiles every 20 minutes without needing any user intervention. For the cases where the user does not have a Google or a Facebook account (or both) or prefers not to log in, the add-on does not collect the attributes from the user's account(s).

Survey: Once the add-on collects the user's attributes from the Google and Facebook accounts for the first time, it displays a survey with up to 100 attributes, selecting randomly 50 of them from each platform, without informing the user about their source, so they can not know if it comes from their Google or Facebook profiles. In this survey, the users are requested to rank each attribute on a scale from 1 (very inaccurate) to 5 (very accurate). The gender and age group attributes (obtained from the user's Google account) are always presented in the survey to be classified based on a binary decision (correct or incorrect) instead of a rank. When the user finishes the classification of the initial 100 attributes, the add-on requests the user to continue with the classification process by clicking on the add-on icon. If the user voluntarily decides to do so, the add-on shows a new list of up to 100 attributes. After each click on the icon, the add-on will display a new list until there are no more attributes to classify. We decided to use this threshold of 100 attributes because it represents a task that can be solved in less than 3 minutes, and represents a minor effort to users that can perform the task without losing focus on it.

Collection of ads: Our add-on is instrumented to identify the presence of Google ads in general websites and Facebook ads in its social network. The add-on collects the following relevant information from each ad: the ad's landing page, the advertising company that distributes the ad, and the ad explanation from Facebook and Google's ad transparency tools. Ads are embedded differently on general websites and the Facebook social network so the add-on implements a different technique for each case.

General Websites: Ads are typically embedded in iFrames [71] inside the HTML of websites. To obtain the ads and all the information associated with them, the add-on follows these steps. First, it discards all iFrames with a size smaller than 20x20 pixels, since they are not big enough for placing ads. Then, for the rest of the iFrames, the add-on searches for all the links (URLs) embedded in the iframe, mostly from the HTML tag `a`. From all the links, the add-on selects those whose Fully Qualified Domain Name (FQDN) belong to an Alphabet company (Google parental company), *e.g.,* `doubleclick.org`. We obtain the list of Alphabet-related companies from the EasyList blocklist [84]. Finally, the add-on retrieves the landing page by searching for specific keywords (`clk=` on the URL) on the ad's URL. It also retrieves the ad explanation if there is a link starting with `https://adssettings.google.com/whythisad?` associated with the ad.

Facebook Social Network: On Facebook, the ads appear as posts on the user's wall. The add-on inspects the requests generated by Facebook while the user is using the social network, selecting only those that contain the path `api/graphql`

in the URL. These requests contain information about the ads received by the user. To retrieve the advertising company and the ad explanation associated with the ad, the add-on identifies the ad id from the referred request. Then, it generates a new request including the ad id and relevant information about the ad to a specific URL[1], obtaining the advertising company and the explanation information of the ad in the response message.

We run a manual experiment to validate the performance of our add-on to collect ads on general websites and Facebook. For the case of websites, we visit 4 times eight popular weather websites[2] that embed a large number of ads from Google. The add-on was able to identify 88% of the ads shown by Google on these pages, over a total of 74 unique ads. For the case of Facebook, we logged to Facebook with the account of one of the authors and navigated through the Facebook newsfeed to guarantee that several ads appeared on the wall. As in the previous case, we repeated the process 4 times, obtaining 80% of the ads shown on the wall.

5.2.2 *Data processing*

Once we implemented and users install the extension on their browsers, we start to collecting the data described in the previous section (see Section 5.2). This data is processes following the the

Attributes Classification: The attributes that appear on the users' profiles can reveal different information about them. Hence, we classified the attributes into three main groups: *general interests*, *sensitive attributes*, and *socio-demographic attributes*. We conducted this classification manually according to the following definitions. First, we identify as *sensitive attributes* those that match with the definition of sensitive interests provided by the EU GDPR in its Article 9 [42]. These attributes may reveal information related to the racial or ethnic origin of the users, their political opinions, their religion or philosophical beliefs, trade union membership, and their sex life or sexual orientation. Second, we classify as *socio-demographic attributes* those that reveal social and demographic aspects of the users. These include basic demographic information (gender and age) as well as other characteristics such as the educational level (*e.g.,* Bachelor's Degree), the home-ownership status (*e.g.,* Renters), the parental status (*e.g.,* Parents of Infants), the marital status (*e.g.,* Married), and the work status (*e.g.,* Job Industry: Healthcare Industry) of the users. Finally, we consider as *general interests* any other attribute that does not match the above definitions.

Attributes popularity: The Facebook Marketing API offers the Monthly Active Users (MAU) and Daily Active Users (DAU) associated with each attribute used by Facebook to profile users. In essence, MAU and DAU are proxy metrics to the number of users that have been assigned a given attribute by the Facebook profiling algorithm. For every Facebook attribute in our dataset, we collect the

1 https://m.facebook.com/nt/screen/?params=<Ad_Attributes>

2 weather.com, wunderground.com, accuweather.com, meteoblue.com, windfinder.com, foreca.com, weatherspark.com, and theweathernetwork.com

estimated worldwide MAU. Unfortunately, Google does not offer information about the popularity of individual attributes.

Profile and ads' attributes mapping: One of our goals is to assess the accuracy of ad targeting algorithms, *i.e., are ad campaigns reaching users interested in their targeted attributes?*. To this end, we map the targeting attributes used by ads with those attributes ranked by the users in the survey. We next explain how we conduct this mapping of profile and ads' attributes on Facebook and Google.

Facebook: The explanation provided by Facebook's, *Why Am I seeing this ad?* for each ad includes a detailed list of attributes used to target users by the associated ad campaign. In addition, Facebook defines a unique dictionary of attributes for users' profiles and for defining targeted ad campaigns. Therefore, there is no need to implement any mapping function. Indeed, it is straightforward to check if an attribute in a targeted ad is present in the targeted user's profile and, if so, what is the score assigned by the user to such attribute through the survey response.

Google: Google's ad explanation is rather poor and does not provide information about the specific attributes used to target a user. Therefore, we have to define a methodology to first infer the targeting attributes associated with an ad and then map the inferred attributes to the user's profile attributes. To assign targeting attributes to an ad, we follow the methodology described in previous works [75]. We identify the landing page of the ad and extract the categories of the landing page using a domain classification service. We decided to use Webshrinker [145], a marketing-oriented domain classification service, which provides labels from the IAB taxonomy [147]. We use Webshrinker as there is not any specific service using Google's taxonomy of categories. IAB [3] is an advertising organization responsible for developing some of the most relevant industry standards and its taxonomy is largely used in online advertising. This process allows us to obtain the ad's targeting attributes. Note that Webshrinker also offers high coverage, defined as the number of websites for which it provides a meaningful category. We based our decision on the analysis of domain classification services presented on Chapter4.

To proceed to the mapping of user's profile attributes and ads' attributes, we manually mapped every attribute of Google's profiles in our dataset to its corresponding category in the IAB taxonomy. At this point, we can check if an ad's targeting attribute is present in the targeted end user's profile. If so, we can retrieve the score the user provided to that attribute in a survey answer.

5.2.3 *Metrics*

Our goal is twofold. First, we want to assess the accuracy of users' profiles generated by Google and Facebook. To this end, we have defined the *Profiling Accuracy* metric as the distribution of scores (on the scale of 1 to 5) given by users to their profile attributes through the survey functionality of our add-on. Second, we want to analyze the actual accuracy of targeted ads by checking if

Platform	Ads			Explanations		
	Total	Unique	Websites	Total	% Ads Explanation	
Google	31	52,721	20,074	1,701	18,297	35%
Facebook	37	141,121	5,047	-	127,125	90%
Unique	49	193,842	25,121	-	145,422	75%

Table 5.2: Description of the dataset. Number ads obtained from General websites (Google) and Facebook.

Platform	Users	Interests				Survey				
		Unique	Users	Unique Sensitive	Unique Soc. Demo.	Responses	Unique Attributes	Users	Sensitive Attributes	Socio. Demo. Attributes
Google	58	1,581	50	7	75	3,256	823	37	4	54
Facebook	51	2,849	33	126	6	3,144	1,644	20	59	0
Unique	62	4,311	57	131	79	6,400	2,409	39	60	54

Table 5.3: Description of the dataset. Number users and Interests obtained form the profiles and ranked by the users.

the users' receiving an ad are interested in the targeted attribute by such ad. We have defined the *Targeting Accuracy* metric for this purpose as the distribution of scores (on the scale of 1 to 5) of attributes of targeted ads. Note that to compute the targeting accuracy we leverage the mapping of profile's attributes and ads' attributes described above.

5.3 DATASET

Before getting into details of the results obtained, we describe the dataset collected with our browser ad-on and processed with the methodology described in the previous section. We summarizes on Table 5.3 the main statistics of the total number of users who installed the add-on and the total number of interests collected and ranked by the users. Table 5.2 summarizes the ads obtained from general websites and Facebook.

Users: Our add-on was installed by 62 users. From 47 of them, we collected data from both Facebook and Google. Moreover, 4 and 11 users only provided data about Facebook or Google, respectively. This paper reports results based on more than 8 months of data collection, from the 1st of January 2021 to the 14th of September 2021. We could collect the profile attributes of all the users. However, not all the users contributed with survey responses or ads to our dataset. In particular, from 92% of users who provided survey responses the add-on collected ads from 80% of them. The ages of the users in our dataset range from 18 to 64. However, our dataset presents a clear bias in terms of gender, since 90% of the users who reported their gender through the survey indicated they are men. While the number of users may seem small, the multiplicative effect of the information collected from these users provides us with sufficient data

Platform	Group	# Interests	Category Example
Facebook	Politics	41	`Left-wing politics`
	Health	29	`Wellness (alternative medicine)`
	Philoshopical Beliefs	23	`Fanatics`
	Religious	21	`Judaism`
	Ethnic	6	`Hispanic american culture`
	Trade Union	5	`Trade association`
	Sexual	1	`Homosexuality`
Google	Health	3	`Psychology`
	Politics	2	`Politics`
	Trade Union	1	`Labor & Employment Law`
	Philoshopical Beliefs	1	`Charity & Philanthropy`

Table 5.4: Number of potentially sensitive attributes identified in our dataset for Facebook and Google per GDPR's sensitive category.

to reveal some relevant insights. Each user in our dataset provides on average 344 profile attributes assigned by Google and Facebook. Moreover, we have collected 2,409 unique interest ranked in the survey (164 interests on average per user) and more than 3,100 ads on average per user. Of course, despite this multiplicative effect, we are cautious and do not claim that the obtained results can be generalized to the whole of Google or Facebook's user base.

Profile Attributes: We have collected a total of 1,581 and 2,849 unique profile attributes on Google and Facebook, respectively. The unique number of attributes obtained is larger for Facebook compared to Google because users are assigned a larger number of attributes in their Facebook profile (245 attributes per user profile on average) than on Google (230 attributes per user profile on average). This is consistent with previous research results [172].

Let us explore the statistics of our dataset concerning the different types of attributes defined in Section 5.2.2 (general interests, sensitive attributes, and socio-demographic attributes). As expected the large majority of attributes correspond to users' interests. We observe that just 4.7% and 0.2% of attributes correspond to socio-demographic properties on Google and Facebook, respectively. On the other hand, only 7 attributes are classified as potentially sensitive in the case of Google. However, 87% of Google users in our dataset have at least one of these attributes assigned to their profile. In the case of Facebook, we classify 126 attributes as potentially sensitive, which have been assigned to 55% of Facebook users in our dataset. Specifically, each user is tagged on average with 11 potentially sensitive attributes. Finally, table 5.4 shows the number of sensitive attributes assigned by Facebook to users in our dataset across GDPR categories. We observed 131 different sensitive interests in total [244].

Survey responses: The users of our add-on provided 6,400 survey answers that assigned scores to 2,409 unique attributes. In particular, we obtained 3,256

(3,144) responses providing a score to 823 (1,644) unique attributes on Google (Facebook).

Ads: We collected 141,121 ad impressions associated to 5,047 unique ads shown to the Facebook users in our dataset. In particular, (on average) 3,814 ad impressions were shown to each Facebook user. In the case of Google, we collected 52,721 ad impressions from 20,074 ads. These ads were collected across 1,005 websites and the average number of ad impressions per user was 1,701.

5.4 RESULTS

In this section, we leverage the described methodology and dataset to compute the *Profiling Accuracy* and *Targeting Accuracy* introduced in Section 5.2. In particular, we separately discuss the results on *Profiling Accuracy*, and *Targeting Accuracy* for the different categories in which we have classified the users' attributes: general interests, sensitive attributes, and socio-demographic attributes. Moreover, we separately present the results obtained for Facebook and Google, and after this, we add a comparative discussion of the two platforms.

5.4.1 *Profiling Accuracy*

Proper performance of profiling algorithms would lead to accurate profiles and thus the Profiling Accuracy resulting from the users' scores should be biased towards high values (4 and 5). Instead, a poor performance would lead to a bias of Profiling Accuracy towards low scores (1 and 2).

5.4.1.1 *General Interests*

Google: Google's profiling algorithm shows poor performance. Over 47% of the survey responses correspond to scores 1 and 2. Indeed, 27% of the responses are a 1, being this the most frequent score among the collected survey responses. Instead, less than 14% of the responses correspond to a score of 5, which is the least frequent score.

Facebook: Almost 50% of responses correspond to score values of 1 or 2, from which a majority (34%) is 1. The score value 5 is again the least frequent with less than 14% of survey responses by users. This indicates an even poorer performance of Facebook's profiling algorithm compared to Google's one. We now factor in the popularity of general interests in our analysis of the Profiling Accuracy. Our initial hypothesis is that popular interests (*e.g.,* Football, L.A. Lakers, Paris, etc.) are easier to infer since there are more and stronger digital footprints associated with them compared to unpopular interests (*e.g.,* Heidelberg University, Caribbean cuisine, etc.) with a weaker digital footprint. To assess the correctness of our hypothesis, we have computed the popularity of each Facebook general interest in our dataset as its worldwide Monthly Active Users (MAU) reported by the Facebook Marketing API.[3] Figure 5.2a presents the distribution

3 Note this analysis could not be conducted for Google since Google does not report the popularity of general interests in their platform.

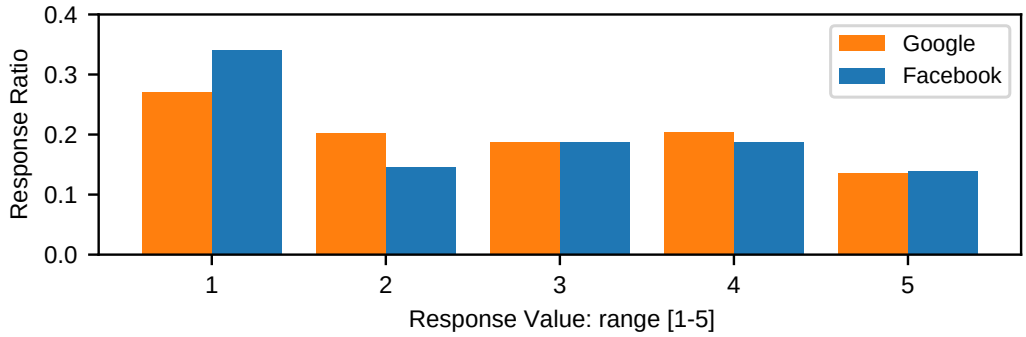
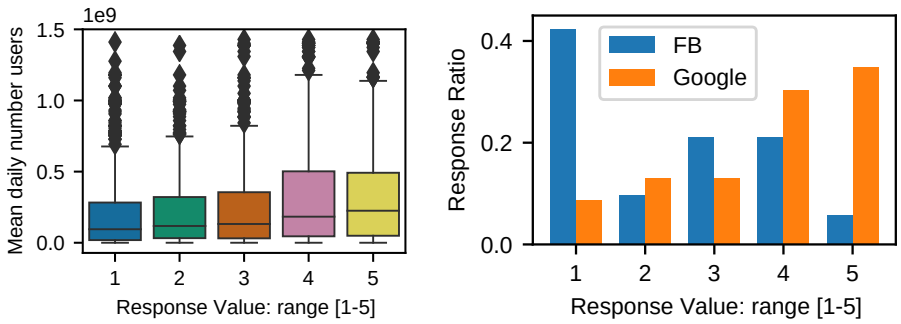


Figure 5.1: Google and Facebook's responses ratio grouped by the rank values.



(a) Facebook interests popularity grouped by the score value assigned in the survey.

(b) Response ratio in survey for potentially sensitive interests grouped by rank value.

Figure 5.2: Facebook interests popularity and Profiling accuracy for sensitive interests on Google and Facebook.

of the popularity of interests ranked with scores 1, 2, 3, 4 and 5 in the form of a box plot. The Y-axis scale reports billions of users worldwide. We observe a clear correlation between the ranking and the popularity, the median popularity of interests ranked with scores (1,2,3,4,5) are (95M, 117M, 131M, 183M, and 225M). This indicates a popularity 2.3 times higher for interests ranked with 5 compared to interests ranked with 1.

Figure 5.1 shows the Profiling Accuracy for the general interests scored by the users of our browser add-on. Each bar represents the percentage of survey responses associated with each of the 5 values in our ranking for Google (Orange bar) and Facebook (Blue bar).

5.4.1.2 *Socio-demographic and sensitive attributes*

In this section, we analyze the Profiling Accuracy for socio-demographic and sensitive attributes.

Google: Table 5.5 shows the distribution of survey responses across the 5 scores of our rank for each one of the socio-demographic categories defined in our classification process. For the case of gender and age group, users use a binary decision.

Due to the specificity of this type of profile's items, we can observe that the responses tend to offer a more binary result, with a major concentration of

	1	2	3	4	5
Education	4	4	5	6	17
Income Level	0	0	1	0	0
House Holder	7	3	3	3	10
Parental Status	16	2	2	5	10
Relationship	9	2	0	3	10
Work	26	6	2	7	10

Table 5.5: Profiling Accuracy for each socio-demographic attribute from Google in our dataset.

responses in the extreme scores 1 and 5, i.e., the profile attribute has been either wrongly or correctly assigned to the user. The results indicate that the Google profiling algorithm seems to offer better performance for socio-demographic attributes than for general interests. Indeed more than 50% of the responses correspond to scores 4 or 5 for every category, except for *Work* that seems to be the hardest one to profile. Finally, we observe that Google tends to assign age and gender correctly. Only 10% and 3% of the users have reported an incorrectly assigned gender and age, respectively. The high accuracy of these two profile items is because they are self-reported by users when they create the accounts.

On the other hand, the 4 sensitive attributes identified in our dataset have been ranked by at least one user of our browser add-on. From all surveys responses associated to sensitive attributes (9%,13%, 13%, 30%, 35%) are associated to scores values (1,2,3,4,5) as shown in Figure 5.2b. We conclude that potentially sensitive attributes seems to be more accurately profiled than general interests.

Facebook: Figure 5.2b shows the Profiling Accuracy for the potentially sensitive interests identified with our manual classification and reported in Table 5.4. The results indicate that, as corroborated by previous works [72] Facebook assigns potentially sensitive interests to their users. However, we report for the first time on the accuracy of such assignation. In particular 52% of the scores of sensitive interests in our dataset are 1 and 2, whereas only 26% correspond to 4s and 5s. This indicates the poor performance of the Facebook profiling algorithm on accurately inferring potentially sensitive interests. On the other hand, our dataset only includes 6 socio-demographic attributes from Facebook. Unfortunately, none of the users scored these attributes in the survey.

5.4.1.3 *Comparison Google vs. Facebook*

The main outcome of our analysis is that both Google and Facebook's profiling algorithms have poor performance. Further relevant insights when we look to the obtained results are the following ones:

1) The distribution of survey responses across the top rank values (3 to 5) is almost identical in the case of Google and Facebook. The difference occurs in the low-rank values. Facebook presents roughly 10 percentage points more

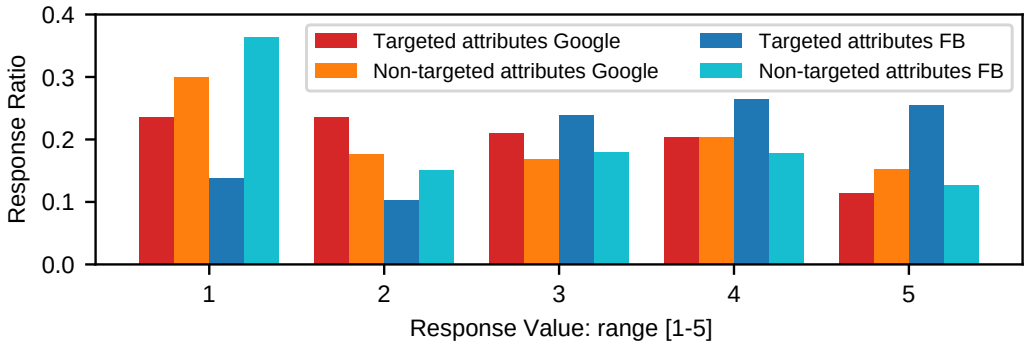


Figure 5.3: Distribution of scores for targeted attributes vs. non-targeted attributes for Google and Facebook in our dataset.

responses in the rank value 1 than Google. If we consider carefully the results, we observe that users have on average a larger number of profile attributes on Facebook than on Google. Moreover, Facebook uses much more specific interests than Google in its profiles, which as shown by Figure 5.2a have a larger probability to fall in low-rank values. In summary, these findings suggest that the excedent of (specific and unpopular) profile attributes used by the Facebook profiling algorithm compared to Google seem to end up in the rank value 1. This would indicate that trying to *overkill* the inference of very specific attributes from users is very hard and leads to less accurate profiles.

2) Our analysis of sensitive interests reveals that Facebook uses in practice a significantly larger number (126) of sensitive interests than Google (7) to profile users. Moreover, our results show that both companies abuse the use of sensitive interests. Google and Facebook have assigned at least one potentially sensitive interest to 87% and 55% of the users in our dataset, respectively. Finally, the analysis of survey responses associated with sensitive interests on Facebook suggests that Facebook offers a poor performance assigning sensitive interests to users of our browser add-on. In conclusion, our results indicate that a major fraction of users expose potentially sensitive data to thousands of advertisers through their Google and Facebook profiles, that in many cases, based on the survey responses for Facebook, is inaccurate sensitive information.

5.4.2 *Targeting Accuracy*

The Targeting Accuracy is a proxy metric that evaluates whether targeted ad campaigns are reaching the right users and thus advertisers receive the service they are paying for. We have computed the Targeting Accuracy using the methodology described in Section 5.2. As in the previous subsection, we organize the presentation of results by first discussing separately them for Facebook and Google, to later introduce a comparative analysis.

5.4.2.1 *General Interests*

We now present the Targeting Accuracy results for the general interests.

Google: Figure 5.3 shows the *Targeting Accuracy* computed from Google's ads in our dataset (red bars) using the ad targeting attributes to user's profile attributes matching technique described in Section 5.2. To give context to these results, the figure also shows the distribution of scores across those attributes ranked by users that are not used in targeted ads in our dataset (orange bars). We refer to these two types of attributes as *targeted* vs. *non-targeted* attributes, respectively. The obtained results provide, to the best of the authors' knowledge, the first evidence of the rather poor performance of Google's targeted ad campaigns. Just 11% of the delivered ads target attributes that users have ranked received a 5 in our surveys, whereas, roughly 22% of the targeted attributes fall in each of the ranks 1 to 4. Based on these results, we conclude that just 1/3 of the ads target meaningful interests for users (*e.g.*, their associated score is 4 or 5). If we compare the distribution of targeted vs. non-targeted attributes across scores, we observe that the overall fraction of meaningful interests (score values 4 or 5) is very similar for both targeted (32%) and non-targeted ads (35%). This seems to indicate that Google does not apply any specific verification algorithm to guarantee that targeted ads impact users that actually meet the targeting criteria of the ad campaign.

Facebook: Figure 5.3 shows the distribution of scores for *targeted* (i.e., the Targeting Accuracy, blue bars) vs. *non-targeted* interests (for context, cyan bars). The results show that Facebook's ad targeting algorithms seem to be significantly more accurate than those used for profiling users. In particular, around 51% and 30% of targeted and non-targeted attributes are ranked with 4 or 5 by the participants, respectively. Instead, 23% of the targeted and 50% non-targeted attributes are ranked with 1 or 2. This provides initial evidence that Facebook has an internal classification algorithm able to rank the accuracy of users' attributes so that the ones used in targeted ad campaigns tend to be those more meaningful for users. In addition, we computed the popularity of targeted and non-targeted attributes. We observe that the average worldwide MAU of targeted and non-targeted attributes is 343M and 135M, respectively. This highlights that the ad campaigns tend to use more popular attributes, and the use of attributes with a lower audience is less frequent. Despite this, and without a clear reason, Facebook seems to assign to users profile attributes infrequently used in targeted advertising campaigns.

5.4.2.2 *Socio-demographic and Sensitive Attributes*

In this subsection, we show the Targeting Accuracy results for sensitive and socio-demographic interests of the users.

Google: Due to the lack of details of the ad transparency tool of Google, we cannot identify if a targeted ad is using sensitive and socio-demographic data from the users' profile. In addition, our normalization methodology is based on a domain classification service that leverages an IAB taxonomy that does not include sensitive and socio-demographic categories.

Facebook: We analyzed the presence of those users' profile attributes classified as potentially sensitive in the list of interests provided by the *Why Am I Seeing This Ad?* for ads in our dataset. Our analysis reveals that 539 ads (24

unique ones) in our dataset use potentially sensitive interests to target users. These ads cover 4 of the 7 sensitive categories defined by the GDPR as shown in Table 5.3. In particular, {443 ads (10 unique ones), 85 (4), 2 (1), 1 (1)} are related to {Health, Politics, Ethics,Trade Union} categories, respectively. Of these sensitive interests used to deliver ads, 7 have been ranked in the survey by 6 different users, obtaining 14 responses. 3 responses were ranked with value 1, 1 with value 2, 8 with value 3, and finally 2 with value 4. This represents preliminary insights on the lack of accuracy to target people based on sensitive attributes on Facebook ads campaigns. Further experiments need to be conducted to achieve significant conclusions.

5.4.2.3 *Comparison of Targeting Accuracy Google vs. Facebook*

The main takeaway from our analysis is that Facebook presents a much better performance to target users with meaningful ads compared to Google. However, this opens a question on Facebook incentives to assign a large number of meaningless interests (*i.e.,* ranked with a low value by participants) to users' profiles. The second important outcome is that the Facebook transparency tool is significantly more *transparent* than Google's. Taking advantage of the data gathered from the Facebook transparency tool we could confirm that advertisers indeed use fairly frequently sensitive interests to target users with Facebook advertising campaigns. The vagueness of Google's transparency tool prevents us from conducting such type of analysis in the case of Google ads.

Part V

ETHICAL CONSIDERATIONS, CONCLUSIONS AND FUTURE WORK

6

ETHICAL CONSIDERATIONS

In this section, we discuss the measurements and actions we took to minimize the impact of our research work from a privacy point of view, complying with the existing legal frameworks and following ethical principles on research, putting particular emphasis on Chapters 3 and 5.

(i) It is essential to remark that the work presented in Chapters 3 does not involve human subjects. All the experiments were run automatically in a controlled environment using crawlers. The processes involving manual inspection were conducted by the members implicated in the work who gave their approval, aware of the possibility of having to see potentially uncomfortable and sensitive images in some cases. Also, before running any experiments using the VPNs, we contacted NordVPN and PrivateVPN to inform them about our research work to ensure we do not break their terms of use and do not harm real-user during the experiments. Furthermore, we do not interact with the consent notices displayed by the websites and do not surf beyond the landing page to avoid generating advertising revenues and accessing specific content. Also, we do not discard the presence of additional tracking mechanisms and services beyond the landing page, as previous research work has demonstrated [245]. Finally, before performing our data collection, we also defined a protocol to report any service distributing illegal pornographic content to the authorities in case this uncomfortable situation arose. Unfortunately, we found one service distributing such content while performing our sanitization process. We immediately reported the case to the national authorities.

(ii) The work presented in Chapter 5 involves collecting data from real users who voluntarily installed the add-on. Before we made the add-on extension public, we asked for approval from the Institutional Ethics Board (IRB) and Data Protection Officer (DPO) of IMDEA Networks Institute. Users provide explicit consent to be part of the study, following the principles of informed consent [246]. We inform them about the privacy risks they face, like disclosing sensitive attributes, before and after the installation. The add-on does not send any data to our server from the users who do not provide explicit consent. Also, the add-on does not collect any personal identifier. Instead, the add-on randomly creates a hexadecimal ID of 256 bytes after the installation to distinguish each user. This ID remains in the browser until it removes it, so all the users that remove and re-install the extension will have a different ID.

7

CONCLUSIONS

The fast growth of the web in size and diversity and the current trend of implementing new regulatory frameworks to limit the collection and processing of personal data required a continuous adaption of fundamental methodologies and technologies to audit and scrutinize web services. This dissertation wants to fill this gap by proposing new methods and measurements to increase the research community, policymakers, and regulators' knowledge about the web ecosystem.

Privacy Analysis of the Web Porn Ecosystem. This need is even more notorious in the online porn ecosystem, which has been traditionally considered an obscure subsystem of the Internet. Nevertheless, the online porn industry is not different from regular web services: it has rapidly integrated advanced tracking technologies to monitor (and, in some cases, monetize) users. However, the third-parties services providing advertising and tracking services to online porn websites differ substantially from those operating on regular websites, even creating a parallel ecosystem concerning regular websites (Section 3.2.2). Moreover, the presence of porn-specific trackers might render many anti-tracking technologies based on blocklists insufficient, as 91% of the scripts implementing canvas fingerprinting do not appear on EasyList and EasyPrivacy lists.

Furthermore, we observe that several porn websites fail to implement common security mechanisms like the use of HTTPS, and basic transparency requirements such as privacy policies and cookie consent forms, even in those websites actively tracking users. Only the companies behind some of the most popular pornographic websites seem to make efforts to comply with current legislation, possibly fearing the high fines of new regulations like the GDPR. Besides data protection and users' privacy, we demonstrated that the efforts made by the online porn industry to prevent children's access to inappropriate content are not being widely deployed. While most countries do not have laws to prevent children from accessing pornographic material, the deployment of these mechanisms is rare even in jurisdictions that want to implement such laws, for instance, the UK.

Remarkably, this contribution opens up new avenues for other studies focused on measuring and characterizing the privacy risks of semi-decoupled and highly sensitive web subsystems (*e.g.,* gambling and online health services) while also informing the public debate. Unfortunately, many of these services might fall between the cracks of public scrutiny and research efforts that aim at identifying web privacy problems from a macroscopic perspective.

An Analysis of Domain Classification Services. Domain classification services have applications in multiple areas, including cybersecurity, online advertising, and academic studies. However, despite their importance in research examinations and certain applications, no previous works have looked at their potential effects.

We find that commonly used domain classification services exhibit traits that affect their suitability. First, only a few services attain a sufficient coverage level to cover non-popular or non-base domains. Second, services may return multiple or undocumented labels, requiring careful data processing and manual validation. Breaking down multi-labeled classification may ease the label comparison between services and improve the interpretation of the results. However, it may also bias the results, overestimating the presence of labels that do not provide information about the real purpose of the service. Third, the large diversity in labels within and across services may harm their accurate and tractable interpretation. Efforts to combine labels from multiple services to achieve a higher agreement on label accuracy might be thwarted by labeling inconsistencies. Finally, the labeling updates may also impact accuracy and timeliness. Researchers should be aware of these phenomena and renew their dataset to reduce possible misclassifications, especially in treating malicious services. In summary, sound deployment and usage of domain classification services require a thorough understanding of the (desired) characteristics and resulting biases to select the most appropriate sources.

We also notice that choosing one domain classification service over another for specialized use cases can significantly impact the accuracy of the results and, consequently, affect academic studies and the effectiveness of solutions relying on them. Yet, we show that human-label services suffer from potential disagreements and could introduce biases. Moreover, labeling is a non-trivial job, as demonstrated by running a small-scale manual classification experiment.

Auditing Profiling and Ad Targeting Algorithms. We present a novel methodology (Chapter 5) to audit the performance of the profiling and ad targeting algorithms from Google and Facebook. Our results show a worrying poor performance of profiling algorithms from both Facebook and Google. We have seen how half of the interests assigned to the profiles by both platforms do not represent the users' online behavior. This poor performance is especially worrying in the case of extensively used sensitive interests on Facebook, whereas previously observed, half of the sensitive ones are not correctly assigned. Regarding the ad targeting algorithms analysis, we see a difference between Google and Facebook. While for the first case, 47% of targeting attributes used to deliver ads are not representative, in the case of Facebook, we see a better ad targeting accuracy with 52% of the targeting attributes being very accurate of the users' profile.

While we openly acknowledge the scale limitations of this study, our results provide initial evidence suggesting that Google should consider an in-depth review of its profiling and ad targeting algorithms. Instead, Facebook seems to have more accurate algorithms that we only applied for ad targeting purposes.

Finally, Facebook should seriously reconsider its frequent use of sensitive interests, given the experienced difficulty in correctly inferring these types of attributes.

8

FUTURE WORK

In the following section, we comment on the main research questions this dissertation has opened and which we plan to address in the near future. As previously, we organize this section based on each of the contributions presented in this dissertation.

Privacy Analysis of the Web Porn Ecosystem. The work presented in Chapter 3 analyzes basic aspects of GDPR compliance. This analysis could be extended, for instance, by analyzing the values of the cookies installed in the browser to investigate the prevalence of other tracking IDs. To extend the analysis of more complex aspects of GDPR compliance, we will need to develop methods to reduce human intervention and supervision, for example, in analyzing the content of privacy policies or the classification of cookie consent banners. Additionally, despite the UK's Digital Economy Act. was not implemented, new efforts are taking place in the UK to impose more restrictive legislation to limit access to minors to porn content. This legislation is included in the Online Safety Bill [54, 55], which is expected to be introduced in 2022. Similar efforts are taking place in European countries like France, where the authorities plan to restrict access to porn websites to minors [56, 57]. We plan to study how these legislative changes could affect porn websites and the suitability of the age verification mechanisms. Likewise, we would like to explore the privacy implications of those websites offering subscription plans by looking at the type of data they require to create the account and comparing the presence and amount of tracking services between the subscription plans and free ones.

An interesting aspect of studying and not addressed in this dissertation is characterizing cross-border data exchanges following the method and techniques reported by Iordanou *et al.* [247] and Razaghpanah *et al.* [95], or by performing a deeper investigation of the connections between online trackers, advertising services, and data brokers.

Finally, we have intentionally not studied aspects such as censorship of pornographic websites and the performance of anti-tracking technologies to protect users' privacy, including safe-browsing modes and popular ad-blockers. We believe that analyzing the effectiveness of such tools longitudinally in specific ecosystems deserves a dedicated study on its own.

An Analysis of Domain Classification Services. We see how domain classification services' characteristics can vary significantly and often tend to be unfavorable; we cannot quantify the quality of individual services due to a lack of comprehensive ground truth. We, therefore, avoid putting forward specific guidance on which services end users and researchers should prefer. Instead,

we provide directions for future work that would bring us closer to such an evaluation.

While we have been able to compare labels between services by analyzing their diversity, understanding their semantic agreement would require developing a new taxonomy to which all labels across all services need to be translated, similar to how AVClass [248] automatically annotates malware samples with one semantically-equivalent label generated from multiple antivirus labels. This translation could occur manually, which may be more accurate, but comes at a higher maintenance cost when taxonomies change, or additional service is to be integrated. Alternatively, this taxonomy development could be (partially) automated through methods such as label normalization, heuristics [249], determining related label pairs between services, or a semantic interpretation of existing labels through natural language processing. Anecdotally, we explored the latter method, but it generated a high false positive rate (*e.g.*, *web spam* and *web hosting* could be reported as equivalent).

Beyond case studies, we do not broadly evaluate label *correctness*: even if all services agree on a label, it might still be wrong. An independently developed classifier can serve as a more trustworthy source of labels against which the labels from other services could be compared. The classifier would need to rely on state-of-the-art automated methods, including topic modeling [250] to cope with the large scale of the Internet. Potential sources of ground truth are human-developed directories such as DMOZ and Curlie (as used in previous work [251, 252, 253, 131]) or the categorization of pages on Wikipedia (*idem* [254]). While an automated model may not achieve perfect accuracy, its methods and performance can be disclosed transparently, improving the soundness of research that depends on it and enabling unbiased evaluation.

These steps could result in a classification service that researchers can rely upon to retrieve category labels obtained through a well-documented process and embedded in a vetted taxonomy. In addition, such a service could either translate the set of labels from existing third-party classification services into labels from a custom taxonomy or output the labels from a custom independent classifier. We consider both challenges to be exciting avenues for future work.

Auditing Profiling and Ad Targeting Algorithms. We developed and implemented a novel method to audit the accuracy of profiling and the ad targeting algorithms from Google and Facebook. However, some improvements in the methodology will allow us to overcome research questions that we have not been able to answer.

However, we first need to increase the representativeness of our dataset by boosting the number of active users. We are aware that the limited number of active users we currently have actively using the extension and their diversity in terms of age, or gender, might affect the results and bias the corresponding conclusions. To incentivize its installation and use, we plan to implement a Firefox version and add new functionalities, including the following ones:

- Inform users about the presence of sensitive categories obtained from their profiles and give them the capacity to remove them from their profiles through the extension.
- Provide the stats about the type of ads each user receives, including the information used to target them. This functionality lets users identify why they receive each advertisement in a friendly interface.

Once the new version is ready, we will recruit users from a crowdsourcing marketplace like Amazon Mechanical Turk platform [255]. However, we will need to restrict the access to users with a new account on Google and Facebook, as well as those having AdBlock extensions installed on the browser or even those using anti-tracking browsers like Brave [256]. The data provided by these new users might enrich our dataset, allowing us to extend the results and conclusions, like the analysis of potential biases regarding the gender, location, or educational levels of the users.

BIBLIOGRAPHY

[1] Council of European Union. *General Data Protection Regulation 679/2016*. [Online; access 6-September-2022]. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN.

[2] California State Legislature. *California Consumer Privacy Act*. [Online; access 6-September-2022]. URL: https://www.caprivacy.org/.

[3] IAB. *The Interactive Advertising Bureau*. [Online; access 30-June-2022]. URL: https://www.iab.com/.

[4] IAB. *Internet Advertising Revenue Report*. [Online; access 30-June-2022]. URL: https://s3.amazonaws.com/media.mediapost.com/uploads/InternetAdvertisingRevenueReportApril2021.pdf.

[5] The New York Times. *Cambridge Analytica and Facebook: The Scandal and the Fallout So Far*. [Online; access 30-June-2022]. URL: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

[6] IETF. *HTTP State Management Mechanism*. [Online; access 30-June-2022]. URL: https://datatracker.ietf.org/doc/html/rfc6265.

[7] Steven Englehardt et al. "Cookies that give you away: The surveillance implications of web tracking." In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. 2015, pp. 289–299.

[8] Aaron Cahn, Scott Alfeld, Paul Barford, and Shanmugavelayutham Muthukrishnan. "An empirical study of web cookies." In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. 2016, pp. 891–901.

[9] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. "Detecting and defending against third-party tracking on the web." In: *Proceedings of the USENIX Conference on Networked Systems Design and Implementation Symposium*. NSDI '12. USENIX Association. 2012, pp. 155–168.

[10] Balachander Krishnamurthy and Craig Wills. "Privacy diffusion on the web: a longitudinal perspective." In: *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. 2009, pp. 541–550.

[11] Mozilla. *Same-origin policy*. [Online; access 6-September-2022]. URL: https://developer.mozilla.org/en-US/docs/Web/Security/Same-origin_policy.

[12] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. "Cookie synchronization: Everything you always wanted to know but were afraid to ask." In: *Proceedings of the 28th International Conference on World Wide Web*. WWW '19. 2019, pp. 1432–1442.

[13] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. "The web never forgets: Persistent tracking mechanisms in the wild." In: *Proceedings of the ACM Conference on Computer and Communication Security*. CCS '14. ACM. 2014, pp. 674–689.

[14] Lukasz Olejnik, Tran Minh-Dung, and Claude Castelluccia. "Selling off privacy at auction." In: (2013). working paper or preprint.

[15] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P Markatos. "Exclusive: How the (synced) cookie monster breached my encrypted vpn session." In: *Proceedings of the 11th European Workshop on Systems Security*. 2018, pp. 1–6.

[16] Muhammad Ahmad Bashir and Christo Wilson. "Diffusion of user tracking data in the online advertising ecosystem." In: *Proceedings of the Privacy Enhancing Technologies Symposium*. Vol. 4. PETS 2018. 2018, pp. 85–103.

[17] 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 ACM ASIA Conference on Computer and Communication Security*. ASIACCS '20. ACM. 2020, pp. 222–235.

[18] Samy Kamkar. *Evercookie*. [Online; access 19-September-2022]. URL: https://samy.pl/evercookie/.

[19] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. "Flash cookies and privacy." In: *2010 AAAI Spring Symposium Series*. 2010.

[20] Peter Eckersley. "How unique is your web browser?" In: *Proceedings of the Privacy Enhancing Technologies Symposium*. PETS 2010. 2010, pp. 1–18.

[21] Aleecia M McDonald and Lorrie Faith Cranor. "A survey of the use of adobe flash local shared objects to respawn http cookies." In: vol. 7. HeinOnline, 2011, p. 639.

[22] Keaton Mowery and Hovav Shacham. "Pixel perfect: Fingerprinting canvas in HTML5." In: vol. 2012. W2SP '12. 2012.

[23] Mozilla. *Canvas API*. [Online; access 6-September-2022]. URL: https://developer.mozilla.org/en-US/docs/Web/API/Canvas_API.

[24] Steven Englehardt and Arvind Narayanan. "Online tracking: A 1-million-site measurement and analysis." In: *Proceedings of the ACM Conference on Computer and Communication Security*. CCS '16. ACM. 2016, pp. 1388–1401.

[25] Alejandro Gómez-Boix, Pierre Laperdrix, and Benoit Baudry. "Hiding in the crowd: an analysis of the effectiveness of browser fingerprinting at large scale." In: *Proceedings of the 27th International Conference on World Wide Web*. WWW '18. 2018, pp. 309–318.

[26] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. "Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints." In: *IEEE Symposium on Security and Privacy*. SP '16. IEEE. 2016, pp. 878–894.

[27] Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy. "Fp-stalker: Tracking browser fingerprint evolutions." In: *IEEE Symposium on Security and Privacy*. SP '18. IEEE. 2018, pp. 728–741.

[28] Keaton Mowery, Dillon Bogenreif, Scott Yilek, and Hovav Shacham. "Fingerprinting information in JavaScript implementations." In: vol. 2. W2SP '11 11. 2011.

[29] Mozilla. *Web Audio API*. [Online; access 30-June-2022]. URL: https://developer.mozilla.org/en-US/docs/Web/API/Web_Audio_API.

[30] EUR-Lex. *ePrivacy proposal*. [Online; access 6-September-2022]. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%5C%3A52017PC0010.

[31] Politico. *How Europe's new privacy rules survived years of negotiations, lobbying and drama*. [Online; access 6-September-2022]. URL: https://www.politico.eu/article/europe-privacy-rules-survived-years-of-negotiations-lobbying.

[32] Maximilian Hils, Daniel W Woods, and Rainer Böhme. "Measuring the emergence of consent management on the web." In: *Proceedings of the Internet Measurement Conference*. IMC '20. New York, NY, USA: ACM, 2020, pp. 317–332.

[33] 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: *IEEE Symposium on Security and Privacy*. SP '20. IEEE. 2020, pp. 791–809.

[34] Cristiana Santos, Arianna Rossi, Lorena Sanchez Chamorro, Kerstin Bongard-Blanchy, and Ruba Abu-Salma. "Cookie Banners, What's the Purpose? Analyzing Cookie Banner Text Through a Legal Lens." In: *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*. 2021, pp. 187–194.

[35] Iskander Sanchez-Rola et al. "Can i opt out yet? gdpr and the global illusion of cookie control." In: *Proceedings of the ACM ASIA Conference on Computer and Communication Security*. ASIACCS '19. ACM. 2019, pp. 340–351.

[36] Jannick Sørensen and Sokol Kosta. "Before and after GDPR: The changes in third party presence at public and private European websites." In: *Proceedings of the 29th International Conference on World Wide Web*. WWW '19. 2019, pp. 1590–1600.

[37] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. "We Value Your Privacy... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy." In: *Proceedings of the Network and Distributed System Security Symposium*. NDSS 2019. 2019.

[38] Cliqz. *GDPR - What happened?* [Online; access 6-September-2022]. URL: https://whotracks.me/blog/gdpr-what-happened.html.

[39] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. "Measuring cookies and web privacy in a post-gdpr world." In: *Proceedings of the International Conference on Passive and Active Network Measurements*. PAM '19. Springer. 2019, pp. 258–270.

[40] Timothy Libert, Lucas Graves, and Rasmus Kleis Nielsen. "Changes in third-party content on European News Websites after GDPR." In: (2018).

[41] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. "4 years of EU cookie law: Results and lessons learned." In: *Proceedings of the Privacy Enhancing Technologies Symposium*. Vol. 2. PETS 2019. 2019, pp. 126–145.

[42] European Comission (Algolia). *GDPR Article 9*. [Online; access 6-September-2022]. URL: https://gdpr.algolia.com/gdpr-article-9.

[43] European Comission. *The EU Internet handbook*. [Online; access 6-September-2022]. URL: http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm.

[44] International Association of Privacy Profesionals. *GDPR matchup: Japan's Act on the Protection of Personal Information*. [Online; access 6-September-2022]. URL: https://iapp.org/news/a/gdpr-matchup-japans-act-on-the-protection-of-personal-information/.

[45] Indian Government. *The Personal Data Protection Bill*. [Online; access 6-September-2022]. URL: https://meity.gov.in/writereaddata/files/Personal_Data_Protection_Bill,2018.pdf.

[46] José González Cabañas, Ángel Cuevas, and Rubén Cuevas. "Unveiling and quantifying facebook exploitation of sensitive personal data for advertising purposes." In: *Proceedings of the USENIX Security Symposium*. USENIX Security '18. 2018, pp. 479–495.

[47] Gareth Tyson, Yehia Elkhatib, Nishanth Sastry, and Steve Uhlig. "Demystifying porn 2.0: A look into a major adult video streaming website." In: *Proceedings of the Internet Measurement Conference*. IMC '13. New York, NY, USA: ACM, 2013, pp. 417–426.

[48] Gilbert Wondracek, Thorsten Holz, Christian Platzer, Engin Kirda, and Christopher Kruegel. "Is the Internet for Porn? An Insight Into the Online Adult Industry." In: *WEIS*. 2010.

[49] Ibrahim Altaweel, Maximillian Hils, and Chris Jay Hoofnagle. "Privacy on adult websites." In: *Workshop on Technology and Consumer Protection, co-located with the 38th IEEE Symposium on Security and Privacy*. ConPro '17. 2016.

[50] Florencia Marotta-Wurgler. "Self-regulation and competition in privacy policies." In: *The Journal of Legal Studies* 45.S2 (2016), S13–S39.

[51] Robert A Gomez. "Protecting minors from online pornography without violating the first amendment: Mandating an affirmative choice." In: *SMU Sci. & Tech. L. Rev.* 11 (2007), p. 1.

[52] Legislation.gov.uk. *Digital Economy Act 2017*. [Online; access 6-September-2022]. URL: http://www.legislation.gov.uk/ukpga/2017/30/contents/enacted.

[53] Wired. *Inside the messy collapse of the UK's unworkable porn block*. [Online; access 6-September-2022]. URL: https://www.wired.co.uk/article/uk-porn-ban-digital-economy-act.

[54] BBC. *Porn sites will be legally required to verify users' age*. [Online; access 6-September-2022]. URL: https://www.bbc.com/news/technology-60293057.

[55] The UK Government: for Digital, Culture, Media & Sport. *Draft Online Safety Bill*. [Online; access 6-September-2022]. URL: https://www.gov.uk/government/publications/draft-online-safety-bill.

[56] Daily Mail. *Major porn sites will be BLOCKED unless they bring in measures to ensure users are over 18 under new French rules*. [Online; access 6-September-2022]. URL: https://www.dailymail.co.uk/news/article-10349501/Major-porn-sites-BLOCKED-France-unless-ensure-users-18.html.

[57] French Republic Government. *Decret n 2021-1306 du 7 octobre 2021*. [Online; access 6-September-2022]. URL: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044173388?s=03.

[58] AgeID. *AgeID announced to the industry at European Summit*. [Online; access 6-September-2022]. URL: https://www.ageid.com/press/article/11.

[59] The Independent. *Porn website age verification tool officially announced within UK*. [Archived; 12-July-2022]. URL: https://web.archive.org/web/20220712133302/https://www.independent.co.uk/life-style/porn-age-verification-tool-uk-announcement-pornhub-ageid-adult-content-websites-mindgeek-a8242476.html.

[60] ASACP. *Association of Sites Advocating Child Protection*. [Online; access 6-September-2022]. URL: https://www.asacp.org/.

[61] ASACP. *ASACP Members*. [Online; access 6-September-2022]. URL: https://www.asacp.org/index.php?content=members%5C#top.

[62] Vice. *Russians now need a passport to watch Pornhub*. [Online; access 6-September-2022]. URL: https://news.vice.com/en_us/article/kzgkv3/russians-now-need-a-passport-to-watch-pornhub.

[63] BBC News. *Russia extends porn site ban*. [Online; access 6-September-2022]. URL: https://www.bbc.com/news/technology-37373244.

[64] OpenNet Initiative. *Iraq*. [Online; access 6-September-2022]. URL: https://opennet.net/research/profiles/iraq%5C#footnote24_is5a386.

[65] USA Today. *China creates stern Internet, e-mail rules*. [Online; access 6-September-2022]. URL: https://usatoday30.usatoday.com/tech/news/2002/01/18/china-internet.htm.

[66] International Amnesty. *Uganda's new anti-human rights laws aren't just punishing LGBTI people*. [Online; access 6-September-2022]. URL: https://www.amnesty.org.uk/uganda-anti-homosexual-act-gay-law-free-speech.

[67] Reuters. *Uganda's "kill the gays" bill shelved again*. [Archived; 29-January-2016]. URL: https://web.archive.org/web/20160129130554/https://af.reuters.com/article/topNews/idAFJOE74C0HP20110513.

[68] Philip Levis. "The collateral damage of internet censorship by dns injection." In: *ACM SIGCOMM CCR* 42.3 (2012), pp. 10–1145.

[69] Paul Pearce et al. "Global measurement of {DNS} manipulation." In: *Proceedings of the USENIX Security Symposium*. USENIX Security '17. 2017, pp. 307–323.

[70] Graham Lowe, Patrick Winters, and Michael L Marcus. "The great DNS wall of China." In: *MS, New York University* 21.1 (2007).

[71] W3Schools. *Iframe HTML*. [Online; access 30-June-2022]. URL: https://www.w3schools.com/tags/tag_iframe.asp.

[72] José González Cabañas, Ángel Cuevas, Aritz Arrate, and Rubén Cuevas. "Does Facebook use sensitive data for advertising purposes?" In: *Communications of the ACM* 64.1 (2020), pp. 62–69.

[73] Amit Datta, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings." In: *Proceedings of the Privacy Enhancing Technologies Symposium*. Vol. 1. PETS 2015. 2015, pp. 92–112.

[74] Athanasios Andreou, Márcio Silva, Fabrıécio Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. "Measuring the Facebook advertising ecosystem." In: *Proceedings of the Network and Distributed System Security Symposium*. NDSS 2019. 2019.

[75] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. "I always feel like somebody's watching me: measuring online behavioural advertising." In: *Proceedings of the International Conference on Emerging Networking EXperiments and Technologies*. CoNEXT '15. 2015, pp. 1–13.

[76] 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*. WWW '14. 2014, pp. 597–608.

[77] Paul L Vasey and Miranda Abild. *A billion wicked thoughts: What the Internet tells us about sexual relationships*. 2013.

[78] *Alexa Websites Ranking*. [Online; access 23-April-2022]. 2022. URL: https://www.alexa.com/topsites/.

[79] Pornhub. *2018 Year in Review*. [Online; access 6-September-2022]. URL: https://www.pornhub.com/insights/2018-year-in-review%5C#2018.

[80] Luxembourg times. *Porn empire reports half billion dollars in revenue – but ends year with loss*. [Online; access 6-September-2022]. URL: https://luxtimes.lu/luxembourg/33248-porn-empire-reports-half-billion-dollars-in-revenue-but-ends-year-with-loss.

[81] The Guardian. *Gay relationships are still criminalised in 72 countries, report finds*. [Online; access 6-September-2022]. URL: https://www.theguardian.com/world/2017/jul/27/gay-relationships-still-criminalised-countries-report.

[82] Google. *Google Policies Help - Adult Content*. [Online; access 6-September-2022]. URL: https://support.google.com/adspolicy/answer/6023699?hl=en.

[83] IAB and PWC. *The Official xHamster 2019 Trend Report*. [Online; access 6-September-2022]. URL: https://xhamster.com/blog/posts/911001.

[84] EasyList. *EasyList*. [Online; access 6-September-2022]. URL: https://easylist.to.

[85] mypornbible. *My Porn Bible*. [Online; access 6-September-2022]. URL: https://mypornbible.com/.

[86] toppornsites. *Top Porn Sites*. [Online; access 6-September-2022]. URL: http://toppornsites.com/.

[87] only4adults. *Only4 Adults*. [Online; access 6-September-2022]. URL: http://only4adults.com.

[88] Top websites. Adult catogory. *Alexa*. [Archived; 22-April-2017]. URL: https://web.archive.org/web/20190422235428/https://www.alexa.com/topsites/category/Top/Adult.

[89] Quirin Scheitle et al. "A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists." In: *Proceedings of the Internet Measurement Conference*. IMC '18. New York, NY, USA: ACM, 2018, pp. 478–493.

[90] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. "Tracking personal identifiers across the web." In: *Proceedings of the International Conference on Passive and Active Network Measurements*. PAM '16. Springer. 2016, pp. 30–41.

[91] NordVPN. *NordVPN*. [Online; access 6-September-2022]. URL: https://nordvpn.com.

[92] PrivateVPN. *Unlock Content and Stay Protected With a Private VPN*. [Online; access 6-September-2022]. 2019. URL: https://privatevpn.com/.

[93] Mohammad Taha Khan, Joe DeBlasio, Geoffrey M Voelker, Alex C Snoeren, Chris Kanich, and Narseo Vallina-Rodriguez. "An empirical analysis of the commercial vpn ecosystem." In: *Proceedings of the Internet Measurement Conference*. IMC '18. New York, NY, USA: ACM, 2018, pp. 443–456.

[94] Allison McDonald et al. "403 Forbidden: A Global View of CDN Geoblocking." In: *Proceedings of the Internet Measurement Conference*. IMC '18. New York, NY, USA: ACM, 2018, pp. 218–230.

[95] Abbas Razaghpanah et al. "Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem." In: (2018).

[96] Thomas Roelleke and Jun Wang. "TF-IDF Uncovered: A Study of Theories and Probabilities." In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore, Singapore: ACM, 2008, pp. 435–442.

[97] The Next Web. *The (almost) invisible men and women behind the world's largest porn sites*. [Online; access 6-September-2022]. URL: %5Cnewline%7Bhttps://thenextweb.com/insider/2016/03/03/the-almost-invisible-men-and-women-behind-the-worlds-largest-porn-sites/%7D.

[98] Futurism. *Data From British Porn Viewers Might Be In The Hands of One Company*. [Online; access 6-September-2022]. URL: https://futurism.com/mindgeek-monopoly-uk-porn-viewers-data.

[99] New York Magazine. *The Geek-Kings of Smut*. [Archived; 28-June-2017]. URL: https://nymag.com/news/features/70985/index4.html.

[100] V. I. Levenshtein. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." In: *Soviet Physics Doklady* (1966).

[101] AdBlock. *Block Ads. Browse safe.* [Online; access 30-Jun-2022]. URL: https://getadblock.com/.

[102] AdblockPlus. *Surf the web with no annoying ads*. [Online; access 30-Jun-2022]. URL: https://adblockplus.org/.

[103] Zhonghao Yu, Sam Macbeth, Konark Modi, and Josep M Pujol. "Tracking the trackers." In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. 2016, pp. 121–132.

[104] *Disconnect Tracking Protection List.* [Online; access 6-September-2022]. URL: https://github.com/disconnectme/disconnect-tracking-protection.

[105] Addthis. *About AddThis*. [Online; access 6-September-2022]. URL: https://www.addthis.com/about/oracle/.

[106] Bluekai. *Oracle Buys Bluekai*. [Online; access 30-April-2022]. URL: https://www.oracle.com/es/corporate/acquisitions/bluekai/.

[107] Oracle. *Oracle Data Marketplace*. [Online; access 6-September-2022]. URL: https://docs.oracle.com/en/cloud/saas/data-cloud/data-cloud-help-center/Help/AudienceDataMarketplace/AudienceDataMarketplace.html.

[108] Umar Iqbal, Peter Snyder, Shitong Zhu, Benjamin Livshits, Zhiyun Qian, and Zubair Shafiq. "Adgraph: A graph-based approach to ad and tracker blocking." In: *IEEE Symposium on Security and Privacy*. SP '20. IEEE. 2020, pp. 763–776.

[109] ZINGY ADS. *Homepage*. [Online; access 6-September-2022]. URL: http://zingyads.com/.

[110] Adult Force. *Homepage*. [Online; access 6-September-2022]. URL: https://www.adultforce.com/%5C#/.

[111] Doctor Web. *Dr.Web Security Space*. [Online; access 30-June-2022]. URL: https://www.drweb.com/.

[112] ExoClick. *Homepage*. URL: https://www.exoclick.com/.

[113] Juicy Ads. *Homepage*. [Online; access 6-September-2022]. URL: https://www.juicyads.com/.

[114] EroAdvertising. *Homepage*. [Online; access 6-September-2022]. URL: https://www.eroadvertising.com/%5C#!/.

[115] Acxiom. *Acxiom*. [Online; access 6-September-2022]. URL: https://www.acxiom.com.

[116] BBC. *Facebook scandal: Who is selling your personal data?* [Online; access 6-September-2022]. URL: https://www.bbc.com/news/technology-44793247.

[117] Wolfie Christl. *Corporate Surveillance in Everyday Life*. [Online; access 6-September-2022]. URL: %5Cnewline%7Bhttps://crackedlabs.org/en/corporate-surveillance%7D.

[118] Imane Fouad, Nataliia Bielova, Arnaud Legout, and Natasa Sarafijanovic-Djukic. "Tracking the Pixels: Detecting Web Trackers via Analyzing Invisible Pixels." In: *arXiv preprint arXiv:1812.01514* (2018).

[119] MaxMind. *Detect Online Fraud and Locate Online Visitors*. [Online; access 6-September-2022]. URL: https://www.maxmind.com/en/home.

[120] Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. "IP geolocation databases: Unreliable?" In: *ACM SIGCOMM Computer Communication Review* 41.2 (2011), pp. 53–56.

[121] Google Developers. *Cookie Matching*. [Online; access 6-September-2022]. URL: https://developers.google.com/authorized-buyers/rtb/cookie-guide.

[122] IAB and PWC. *WebRTC*. [Online; access 6-September-2022]. URL: https://webrtc.org/.

[123] Philipp Richter et al. "A multi-perspective analysis of carrier-grade NAT deployment." In: *Proceedings of the Internet Measurement Conference*. IMC '16. New York, NY, USA: ACM, 2016, pp. 215–229.

[124] Abbas Razaghpanah et al. "Exploring the design space of longitudinal censorship measurement platforms." In: *arXiv preprint arXiv:1606.01979* (2016).

[125] Tarun Kumar Yadav, Akshat Sinha, Devashish Gosain, Piyush Kumar Sharma, and Sambuddho Chakravarty. "Where The Light Gets In: Analyzing Web Censorship Mechanisms in India." In: *Proceedings of the Internet Measurement Conference*. IMC '18. New York, NY, USA: ACM, 2018, pp. 252–264.

[126] Crypto Webminer. *Crypto Webminer - Web mining - Mining in your Browser*. [Online; access 6-September-2022]. URL: https://www.crypto-webminer.com.

[127] 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: *Proceedings of the USENIX Security Symposium*. 2018, pp. 531–548.

[128] Yahoo! *Yahoo! Directory*. [Archived; 22-November-2014]. URL: https://web.archive.org/web/20141122194515/https://dir.yahoo.com/.

[129] Curlie Project Inc. *Curlie*. [Online; access 30-June-2022]. URL: https://curlie.org/.

[130] AOL Inc. *DMOZ - The Directory of the Web*. [Archived; 14-March-2017]. 2017. URL: https://web.archive.org/web/20170314000301/http://www.dmoz.org/.

[131] Hsin-Chang Yang and Chung-Hong Lee. "A text mining approach on automatic generation of web directories and hierarchies." In: *Expert Systems with Applications* 27.4 (2004), pp. 645–663. DOI: 10.1016/j.eswa.2004.06.009.

[132] Xiaoguang Qi and Brian D. Davison. "Web Page Classification: Features and Algorithms." In: *ACM Computing Surveys* 41.2 (Feb. 2009). DOI: 10.1145/1459352.1459357.

[133] Daniel López-Sánchez, Angélica González Arrieta, and Juan M. Corchado. "Visual content-based web page categorization with deep transfer learning and metric learning." In: *Neurocomputing* 338 (2019), pp. 418–431. DOI: 10.1016/j.neucom.2018.08.086.

[134] Renato Bruni and Gianpiero Bianchi. "Website categorization: A formal approach and robustness analysis in the case of e-commerce detection." In: *Expert Systems with Applications* 142 (2020). DOI: 10.1016/j.eswa.2019.113001.

[135] Chen-Huei Chou, Atish P. Sinha, and Huimin Zhao. "Commercial Internet filters: Perils and opportunities." In: *Decision Support Systems* 48.4 (2010), pp. 521–530. DOI: 10.1016/j.dss.2009.11.002.

[136] Bitdefender. *Bitdefender is the Global Leader in Cybersecurity*. [Online; access 30-June-2022]. URL: https://www.bitdefender.com/.

[137] Symantec Corporation. *Symantec Sitereview*. [Online; access 30-June-2022]. URL: https://sitereview.bluecoat.com/.

[138] McAfee LLC. *Customer URL Ticketing System*. [Online; access 30-June-2022]. URL: https://sitelookup.mcafee.com/.

[139] OpenDNS. *OpenDNS Community: Domain Tagging*. [Online; access 30-June-2022]. URL: https://community.opendns.com/domaintagging/.

[140] Federal Communications Commission. *Children's Internet Protection Act (CIPA)*. [Online; access 30-June-2022]. URL: https://www.fcc.gov/consumers/guides/childrens-internet-protection-act.

[141] R. S. Rosenberg. "Controlling Access to the Internet: The Role of Filtering." In: *Ethics and Information Technology* 3.1 (Mar. 2001), pp. 35–54. DOI: 10.1023/A:1011431908368.

[142] Monica T. Whitty. "Should Filtering Software be utilised in the Workplace? Australian Employees' Attitudes towards Internet usage and Surveillance of the Internet in the Workplace." In: *Surveillance & Society* 2.1 (Sept. 2002). DOI: 10.24908/ss.v2i1.3326.

[143] Paul J. Resnick, Derek L. Hansen, and Caroline R. Richardson. "Calculating Error Rates for Filtering Software." In: *Communications of the ACM* 47.9 (2004), pp. 67–71.

[144] Google. *About contextual targeting*. [Online; access 30-June-2022]. URL: https://support.google.com/google-ads/answer/2404186.

[145] DNSFilter Inc. *Webshrinker*. [Online; access 30-June-2022]. URL: https://www.webshrinker.com/.

[146] Melissa (IAB) Gallo. *Taxonomy: The Most Important Industry Initiative You've Probably Never Heard Of*. [Online; access 30-June-2022]. URL: https://www.iab.com/news/taxonomy-important-industry-initiative-youve-probably-never-heard/.

[147] IAB Tech Lab. *Taxonomy*. [Online; access 30-June-2022]. URL: https://www.iab.com/guidelines/taxonomy/.

[148] Walter Rweyemamu, Tobias Lauinger, Christo Wilson, William Robertson, and Engin Kirda. "Clustering and the Weekend Effect: Recommendations for the Use of Top Domain Lists in Security Research." In: *Proceedings of the International Conference on Passive and Active Network Measurements*. PAM '19. Springer. 2019, pp. 161–177.

[149] Philippe Skolka, Cristian-Alexandru Staicu, and Michael Pradel. "Anything to Hide? Studying Minified and Obfuscated Code in the Web." In: *Proceedings of the 28th International Conference on World Wide Web*. WWW '19. 2019, pp. 1735–1746.

[150] Muhammad Ikram, Rahat Masood, Gareth Tyson, Mohamed Ali Kaafar, Noha Loizon, and Roya Ensafi. "The chain of implicit trust: An analysis of the web third-party resources loading." In: *Proceedings of the 28th International Conference on World Wide Web*. WWW '19. 2019, pp. 2851–2857.

[151] Rebekah Houser, Zhou Li, Chase Cotton, and Haining Wang. "An investigation on information leakage of DNS over TLS." In: *Proceedings of the International Conference on Emerging Networking EXperiments and Technologies*. CoNEXT '19. 2019, pp. 123–137.

[152] Ceren Budak. "What happened? The Spread of Fake News Publisher Content During the 2016 US Presidential Election." In: *Proceedings of the 28th International Conference on World Wide Web*. WWW '19. 2019, pp. 139–150.

[153] Emily Stark et al. "Does certificate transparency break the web? Measuring adoption and error rate." In: *IEEE Symposium on Security and Privacy*. SP '19. IEEE. 2019, pp. 211–226.

[154] Sergio Pastrana, Alice Hutchings, Daniel Thomas, and Juan Tapiador. "Measuring eWhoring." In: *Proceedings of the Internet Measurement Conference*. IMC '19. New York, NY, USA: ACM, 2019, pp. 463–477.

[155] Hang Hu, Peng Peng, and Gang Wang. "Characterizing pixel tracking through the lens of disposable email services." In: *IEEE Symposium on Security and Privacy*. 2019.

[156] Damilola Ibosiola, Ignacio Castro, Gianluca Stringhini, Steve Uhlig, and Gareth Tyson. "Who watches the watchmen: Exploring complaints on the web." In: *Proceedings of the 28th International Conference on World Wide Web*. WWW '19. 2019, pp. 729–738.

[157] Peng Peng, Limin Yang, Linhai Song, and Gang Wang. "Opening the blackbox of virustotal: Analyzing online phishing scan engines." In: *Proceedings of the Internet Measurement Conference*. IMC '19. New York, NY, USA: ACM, 2019, pp. 478–485.

[158] Hiroaki Suzuki, Daiki Chiba, Yoshiro Yoneya, Tatsuya Mori, and Shigeki Goto. "ShamFinder: An Automated Framework for Detecting IDN Homographs." In: *Proceedings of the Internet Measurement Conference*. IMC '19. New York, NY, USA: ACM, 2019, pp. 449–462.

[159] Xianghang Mi et al. "Resident evil: Understanding residential IP proxy as a dark service." In: *IEEE Symposium on Security and Privacy*. 2019.

[160] Matthew Joslin, Neng Li, Shuang Hao, Minhui Xue, and Haojin Zhu. "Measuring and Analyzing Search Engine Poisoning of Linguistic Collisions." In: *IEEE Symposium on Security and Privacy*. 2019.

[161] Victor Le Pochat, Tom Van Goethem, and Wouter Joosen. "Funny accents: Exploring genuine interest in internationalized domain names." In: *Proceedings of the International Conference on Passive and Active Network Measurements*. PAM '19. Springer. 2019, pp. 178–194.

[162] Gong Chen, Wei Meng, and John Copeland. "Revisiting mobile advertising threats with MAdLife." In: *Proceedings of the 28th International Conference on World Wide Web*. WWW '19. 2019, pp. 207–217.

[163] Francesco Marcantoni, Michalis Diamantaris, Sotiris Ioannidis, and Jason Polakis. "A Large-scale Study on the Risks of the HTML5 WebAPI for Mobile Sensor-based Attacks." In: *Proceedings of the 28th International Conference on World Wide Web*. WWW '19. 2019, pp. 3063–3071.

[164] Hsu-Chun Hsiao et al. "An Investigation of Cyber Autonomy on Government Websites." In: *Proceedings of the 28th International Conference on World Wide Web*. WWW '19. 2019, pp. 2814–2821.

[165] SurfControl. *Juniper Test-a-Site*. [Online; access 30-April-2020]. URL: http://mtas.surfcontrol.com/mtas/JuniperTest-a-Site.php.

[166] Computerworld; Mullins, Robert. *Websense makes $400M bid for SurfControl*. [Online; access 30-June-2022]. URL: https://www.computerworld.com/article/2545021/websense-makes--400m-bid-for-surfcontrol.html.

[167] Sazzadur Rahaman, Gang Wang, and Danfeng Yao. "Security Certification in Payment Card Industry: Testbeds, Measurements, and Recommendations." In: *Proceedings of the ACM Conference on Computer and Communication Security*. CCS '19. 2019, pp. 481–498.

[168] Pelayo Vallina, Álvaro Feal, Julien Gamba, Narseo Vallina-Rodriguez, and Antonio Fernández Anta. "Tales from the porn: A comprehensive privacy analysis of the web porn ecosystem." In: *Proceedings of the Internet Measurement Conference*. IMC '19. ACM, 2019, pp. 245–258.

[169] Charles Reis, Alexander Moshchuk, and Nasko Oskov. "Site isolation: process separation for web sites within the browser." In: *Proceedings of the USENIX Security Symposium*. USENIX Security '19. 2019, pp. 1661–1678.

[170] Andrea Morichetta, Martino Trevisan, and Luca Vassio. "Characterizing web pornography consumption from passive measurements." In: *Proceedings of the International Conference on Passive and Active Network Measurements*. Springer. 2019, pp. 304–316.

[171] Santiago Vargas, Utkarsh Goel, Moritz Steiner, and Aruna Balasubramanian. "Characterizing JSON Traffic Patterns on a CDN." In: *Proceedings of the Internet Measurement Conference*. IMC '19. New York, NY, USA: ACM, 2019, pp. 195–201.

[172] Muhammad Ahmad Bashir, Umar Farooq, Maryam Shahid, Muhammad Fareed Zaffar, and Christo Wilson. "Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers." In: *Proceedings of the Network and Distributed System Security Symposium*. NDSS 2019. 2019.

[173] Adam Oest, Yeganeh Safaei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Kevin Tyers. "Phishfarm: A scalable framework for measuring the effectiveness of evasion techniques against browser phishing blacklists." In: *IEEE Symposium on Security and Privacy*. 2019.

[174] Panayiotis Smeros, Carlos Castillo, and Karl Aberer. "SciLens: evaluating the quality of scientific news articles using social media and scientific literature indicators." In: *Proceedings of the 28th International Conference on World Wide Web*. WWW '19. 2019.

[175] Kang-Min Kim, Yeachan Kim, Jungho Lee, Ji-Min Lee, and SangKeun Lee. "From small-scale to large-scale text classification." In: *Proceedings of the 28th International Conference on World Wide Web*. WWW '19. 2019, pp. 853–862.

[176] Brian OpenDNS Community - Idea Bank; Hartvigsen. *Faster Domain Tagging And Approval*. [Online; access 30-June-2022]. URL: https://support.opendns.com/hc/en-us/community/posts/220012547/comments/224509067.

[177] OpenDNS. *FAQ for Domain Tagging*. [Online; access 30-June-2022]. URL: https://community.opendns.com/domaintagging/faq/.

[178] OpenDNS. *OpenDNS Community: Domain Tagging: Categories*. [Online; access 30-June-2022]. URL: https://community.opendns.com/domaintagging/categories.

[179] McAfee Inc. *TrustedSource Web Database, Reference Guide, Category Set 4*. [Online; access 30-June-2022]. URL: https://sitelookup.mcafee.com/download/ts_wd_reference_guide.pdf.

[180] Fortinet Inc. *FortiGuard Labs: Web Filter Lookup*. [Online; access 30-June-2022]. URL: https://fortiguard.com/webfilter.

[181] Fortinet Inc. *FortiGuard Web Filtering*. [Online; access 30-June-2022]. URL: https://docs.fortinet.com/document/fortigate/6.0.0/handbook/120269/fortiguard-web-filtering.

[182] Fortinet Inc. *Web Filter Categories*. [Online; access 30-June-2022]. URL: https://fortiguard.com/webfilter/categories.

[183] VirusTotal. *VirusTotal*. [Online; access 30-June-2022]. URL: http://www.virustotal.com/.

[184] Chronicle Security. *VirusTotal - How it works*. [Online; access 30-June-2022]. URL: https://support.virustotal.com/hc/en-us/articles/115002126889-How-it-works.

[185] Alexa Internet Inc. *Alexa: Top sites by Category*. [Online; access 01-April-2021]. URL: https://www.alexa.com/topsites/category.

[186] Alexa Internet Inc. *Where do Alexa's Top Sites by Category come from?* [Online; access 27-April-2020]. URL: https://support.alexa.com/hc/en-us/articles/200449844.

[187] Alexa Internet Inc. *Is Popularity in the Top Sites by Category directory based on Traffic Rank?* [Online; access 05-May-2020]. URL: https://support.alexa.com/hc/en-us/articles/200461970.

[188] Bitdefender. *Web Filtering Software Development Kit (SDK)*. [Online; access 30-June-2022]. URL: https://download.bitdefender.com/resources/media/materials/2019/pan/en/Bitdefender-OEM-WebFiltering-SDK-Datasheet-creatent169-en_EN-Screen.pdf.

[189] Bitdefender. *Web Categories in GravityZone Content Control*. [Online; access 30-June-2022]. URL: https://www.bitdefender.com/support/web-categories-in-gravityzone-content-control-2287.html.

[190] SiliconANGLE. *Websense acquires Stonesoft from Intel Security, renames combined company Forcepoint*. [Online; access 30-June-2022]. URL: https://siliconangle.com/2016/01/14/raytheon-websense-acquires-stonesoft-from-intel-security-renames-combined-company-forcepoint/.

[191] Forcepoint Inc. *Real-time Threat Analysis with CSI: ACE Insight - Websense.com*. [Online; access 30-June-2022]. URL: https://csi.forcepoint.com/.

[192] Forcepoint. *Master Database URL Categories*. [Online; access 30-June-2022]. URL: https://www.forcepoint.com/product/feature/master-database-url-categories.

[193] Forcepoint. *CSI: ACE Insight Frequently Asked Questions*. [Online; access 30-June-2022]. URL: https://csi.forcepoint.com/Home/Faq.

[194] Doctor Web. *Free Dr.Web online scanner for scanning suspicious files and links*. [Online; access 30-June-2022]. URL: https://vms.drweb.com/online/.

[195] Trend Micro Incorporated. *Trend Micro Site Safety Center*. [Online; access 30-June-2022]. URL: https://global.sitesafety.trendmicro.com/.

[196] Trend Micro. *Web Reputation Services (WRS) Lookup process in Officescan*. [Online; access 30-June-2022]. URL: https://success.trendmicro.com/solution/1056324-web-reputation-services-wrs-lookup-process-in-officescan-osce.

[197] Trend Micro Incorporated. *URL Filtering Categories for Worry Free Business Security Services (WFBS-SVC)*. [Online; access 30-June-2022]. URL: https://success.trendmicro.com/solution/1059905-url-filtering-categories-for-worry-free-business-security-services-wfbs-svc.

[198] Trend Micro. *Website classifications*. [Archived; 11-November-2011]. URL: https://web.archive.org/web/20111111013335/http://solutionfile.trendmicro.com/solutionfile/Consumer/new-web-classification.html.

[199] Trend Micro. *Website classifications*. [Online; access 24-April-2020]. URL: http://solutionfile.trendmicro.com/solutionfile/Consumer/new-web-classification.html.

[200] Symantec. *Symantec WebPulse*. [Online; access 30-June-2022]. URL: https://docs.broadcom.com/doc/webpulse-en.

[201] Symantec. *The Need for Threat Risk Levels in Secure Web Gateways*. [Online; access 30-June-2022]. URL: https://docs.broadcom.com/doc/need-for-threat-risk-levels-in-secure-web-gateways-en.

[202] Symantec Corporation. *Category Descriptions*. [Online; access 30-June-2022]. URL: https://sitereview.bluecoat.com/%5C#/category-descriptions.

[203] Broadcom Inc. *[ALERT] 2019 Symantec Intelligence Services and WebFilter Category and Application Update*. [Online; access 30-June-2022]. URL: https://knowledge.broadcom.com/external/article?articleId=185063.

[204] DNSFilter Inc. *Behind the Curtain of DNSFilter's AI*. [Online; access 30-June-2022]. URL: https://www.dnsfilter.com/wp-content/uploads/2019/04/How_Webshrinker_Works.pdf.

[205] Webshrinker. *APIs - Webshrinker*. [Online; access 30-June-2022]. URL: https://www.webshrinker.com/apis/%5C#domain-category.

[206] Webshrinker. *Website Category API Reference*. [Online; access 30-June-2022]. URL: https://docs.webshrinker.com/v3/website-category-api.html.

[207] Jason Vice Motherboard; Koebler. *AOL Is Mysteriously Shutting Down the 19-Year-Old Community That Inspired Wikipedia*. [Online; access 30-June-2022]. URL: https://www.vice.com/en_us/article/bmbd4m/aol-is-mysteriously-shutting-down-the-19-year-old-community-that-inspired-wikipedia.

[208] AOL Inc. *DMOZ - Suggest a Site: FAQ*. [Archived; 4-March-2017]. URL: https://web.archive.org/web/20170304122239/https://www.dmoz.org/docs/en/help/submit.html.

[209] AOL Inc. *DMOZ - Editorial Guidelines - Subcategories*. [Archived; 3-March-2017]. URL: https://web.archive.org/web/20170303062930/http://www.dmoz.org/docs/en/guidelines/subcategories.html.

[210] Quirin Scheitle et al. "A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists." In: *Proceedings of the Internet Measurement Conference*. IMC '18. New York, NY, USA: ACM, 2018, pp. 478–493.

[211] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation." In: *Proceedings of the Network and Distributed System Security Symposium*. NDSS 2019. 2019.

[212] *Domain Name Industry Brief*. [Online; access 30-June-2022]. Mar. 2020. URL: https://www.verisign.com/assets/domain-name-report-Q42019.pdf.

[213] Google. *Chrome User Experience Report*. [Online; access 30-June-2022]. URL: https://developers.google.com/web/tools/chrome-user-experience-report.

[214] Common Crawl Foundation. *Common Crawl*. [Online; access 30-June-2022]. URL: https://commoncrawl.org/.

[215] OpenDNS. *OpenDNS Community: Domain Tagging: Categories*. [Online; access 30-June-2022]. URL: https://web.archive.org/web/20111108035057/https://community.opendns.com/domaintagging/categories.

[216] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. 2nd. Wiley-Interscience, 2006.

[217] Curlie Project Inc. *Curlie - Editorial Guidelines - Subcategories*. [Online; access 30-June-2022]. URL: https://curlie.org/docs/en/guidelines/subcategories.html.

[218] politicalscienceforias2016. *Homepage*. [Online; access 30-June-2022]. URL: politicalscienceforias2016.wordpress.com.

[219] Black Hawk Colorado Casinos. *Homepage*. [Online; access 30-June-2022]. URL: blackhawkcolorado.com.

[220] Álvaro Feal, Paolo Calciati, Narseo Vallina-Rodriguez, Carmela Troncoso, and Alessandra Gorla. "Angel or Devil? A Privacy Study of Mobile Parental Control Apps." In: vol. 2. PETS 2020. 2020, pp. 314–335.

[221] EasyList. *EasyPrivacy*. [Online; access 6-September-2022]. URL: https://easylist.to/tag/easyprivacy.html.

[222] AdBlockPlus. *Adblock Plus filters explained*. [Online; access 30-June-2022]. URL: https://adblockplus.org/filter-cheatsheet.

[223] Dirección General de Ordenación del Juego. *Listado de URLs de operadores con licencia*. [Online; access 30-June-2022]. URL: https://www.ordenacionjuego.es/es/url-operadores.

[224] Belgian Gaming Commission. *Official list of the Gaming Commission*. [Online; access 30-June-2022]. URL: https://www.gamingcommission.be/opencms/opencms/jhksweb_en/establishments/Online/fplus/.

[225] The official Isle of Man Government Site. *Licence holders*. [Online; access 30-June-2022]. URL: https://www.gov.im/categories/business-and-industries/gambling-and-e-gaming/licence-holders/.

[226] WebPageTest. *optimization_checks.py*. [Online; access 30-June-2022]. URL: https://github.com/WPO-Foundation/wptagent/blob/baab610/internal/optimization_checks.py%5C#L62.

[227] Marjorie Heins, Christina Cho, and Ariel Feldman. *Internet Filters: A Public Policy Report*. Tech. rep. [Online; access 30-June-2022]. Brennan Center for Justice, May 17, 2006. URL: https://www.brennancenter.org/sites/default/files/2019-08/Report_Internet-Filters-2nd-edition.pdf.

[228] Sarah Houghton-Jan. "Internet Filtering." In: *Library Technology Reports* 46.8 (Nov. 2010), pp. 25–33.

[229] Caroline R. Richardson, Paul J. Resnick, Derek L. Hansen, Holly A. Derry, and Victoria J. Rideout. "Does Pornography-Blocking Software Block Access to Health Information on the Internet?" In: *JAMA* 288.22 (Dec. 2002), pp. 2887–2894. DOI: 10.1001/jama.288.22.2887.

[230] Matthew Cloudflare; Prince. *The Mistake that Caused 1.1.1.3 to Block LGBTQIA+ Sites Today*. [Online; access 30-June-2022]. URL: https://blog.cloudflare.com/the-mistake-that-caused-1-1-1-3-to-block-lgbtqia-sites-today/.

[231] Backlinko). *Ad Blocker Usage and Demographic Statistics in 2021*. [Online; access 30-June-2022]. URL: https://backlinko.com/ad-blockers-users.

[232] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. "What Makes a "Bad" Ad? User Perceptions of Problematic Online Advertising." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–24.

[233] Laura Shipp and Jorge Blasco. "How private is your period?: A systematic analysis of menstrual app privacy policies." In: *Proceedings of the Privacy Enhancing Technologies Symposium*. Vol. 4. PETS 2020. 2020, pp. 491–510.

[234] The Wall Street Journal. *Amazon Surpasses 10% of U.S. Digital Ad Market Share*. [Online; access 30-June-2022]. URL: https://www.wsj.com/articles/amazon-surpasses-10-of-u-s-digital-ad-market-share-11617703200.

[235] Google. *Google Ad Settings*. [Online; access 30-June-2022. URL: https://adssettings.google.com/.

[236] Facebook. *Facebook Ad Preferences*. [Online; access 30-June-2022]. URL: https://www.facebook.com/adpreferences/ad_settings.

[237] Google. *About this Ad*. [Online; access 30-June-2022]. URL: https://support.google.com/ads/answer/1634057?hl=en.

[238] Facebook. *Why am I seeing ads from an advertiser on Facebook?* [Online; access 30-June-2022]. URL: https://www.facebook.com/help/794535777607370.

[239] WIRED. *All the Ways Google Tracks You—And How to Stop It*. [Online; access 6-September-2022]. URL: https://www.wired.com/story/google-tracks-you-privacy/.

[240] Google. *Real-time Bidding*. [Online; access 6-September-2022]. URL: https://developers.google.com/authorized-buyers/rtb/downloads/publisher-verticals.

[241] TechCrunch. *Google and IAB ad category lists show 'massive leakage of highly intimate data,' GDPR complaint claims*. [Online; access 6-September-2022]. URL: https://techcrunch.com/2019/01/27/google-and-iab-ad-category-lists-show-massive-leakage-of-highly-intimate-data-gdpr-complaint-claims/.

[242] Jan-Willem van Dam and Michel Van De Velden. "Online profiling and clustering of Facebook users." In: *Decision Support Systems* 70 (2015), pp. 60–72.

[243] Two Wheels Marketing. *Facebook Ad Targeting Options*. [Online; access 6-September-2022]. URL: https://twowheelsmarketing.com/blog/facebook-ads-targeting-options-list/.

[244] You do not know me project. *List of Sensitive Interests*. [Online; access 30-June-2022]. URL: %5Cnewline%7Bhttps://github.com/youdonotknowmeproject/youdonotknowmeproject/blob/main/sensitive%5C_interests.csv%7D.

[245] Waqar Aqeel, Balakrishnan Chandrasekaran, Anja Feldmann, and Bruce M Maggs. "On landing and internal web pages: The strange case of jekyll and hyde in web performance measurement." In: *Proceedings of the Internet Measurement Conference*. IMC '20. New York, NY, USA: ACM, 2020, pp. 680–695.

[246] Erin Kenneally and David Dittrich. "The Menlo Report: Ethical principles guiding information and communication technology research." In: *Available at SSRN 2445102* (2012).

[247] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. "Tracing cross border web tracking." In: *Proceedings of the Internet Measurement Conference*. IMC '18. New York, NY, USA: ACM, 2018, pp. 329–342.

[248] Marcos Sebastián, Richard Rivera, Platon Kotzias, and Juan Caballero. "AVclass: A Tool for Massive Malware Labeling." In: *19th International Symposium on Research in Attacks, Intrusions, and Defenses*. RAID '16. 2016, pp. 230–253. DOI: 10.1007/978-3-319-45719-2_11.

[249] Jung-Hyun Lee, Jongwoo Ha, Jin-Yong Jung, and Sangkeun Lee. "Semantic Contextual Advertising Based on the Open Directory Project." In: *ACM Transactions on the Web* 7.4 (Nov. 2013). DOI: 10.1145/2529995.2529997.

[250] Ben Weinshel et al. "Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing." In: *Proceedings of the ACM Conference on Computer and Communication Security*. CCS '19. 2019, pp. 149–166.

[251] Chandra Chekuri, Michael H. Goldwasser, Prabhakar Raghavan, and Eli Upfal. "Web Search Using Automated Classification." In: *Proceedings of the 6th International Conference on World Wide Web*. WWW '97. 1997.

[252] Dunja Mladenić. "Turning Yahoo into an Automatic Web-Page Classifier." In: *13th European Conference on Artificial Intelligence*. ECAI '98. 1998, pp. 473–474.

[253] Hao Chen and Susan Dumais. "Bringing Order to the Web: Automatically Categorizing Search Results." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '00. 2000, pp. 145–152. DOI: 10.1145/332040.332418.

[254] Eda Baykan, Monika Henzinger, Ludmila Marian, and Ingmar Weber. "A Comprehensive Study of Features and Algorithms for URL-Based Topic Classification." In: *ACM Transactions on the Web* 5.3 (July 2011). DOI: 10.1145/1993053.1993057.

[255] Amazon. *Amazon Mechanical Turk*. [Online; access 13-September-2022]. URL: https://www.mturk.com/.

[256] Brave. *The best privacy online*. [Online; access 13-September-2022]. URL: https://brave.com.