PRIVACY IN ONLINE ADVERTISING PLATFORMS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF TELEMATICS ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF UNIVERSITAT POLITÈCNICA DE CATALUNYA
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

José Antonio Estrada Jiménez
July 2020

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Jordi Forné Muñoz) Principal Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Javier Parra Arnau) Principal Co-Adviser

Approved for the University Committee on Graduate Studies.

A A. A.

Abstract

Online advertising is consistently considered as the pillar of the "free" content on the Web. By giving websites a way of financing their operation, advertising would be preventing users from being charged for the content consumed on the Web. Besides promoting content creation, this revolution of the marketing business created a myriad of new opportunities for advertisers to reach potential customers at the right time.

Furthermore, the option of delivering personalized ads has turned advertising into a service that can be really valuable for end users, who thank receiving ads tailored to their interests. Given its apparent success in getting paying customers, online advertising is fueling a billionaire business in which the largest tech companies are involved.

The current advertising model builds upon an intricate infrastructure composed of a variety of intermediary entities and technologies whose main aim is to deliver personalized ads. For this purpose, a wealth of user data is collected, aggregated, processed and traded behind the scenes at an unprecedented rate. Despite the enormous value of online advertising, however, the intrusiveness and ubiquity of these practices prompt serious privacy concerns.

In view of the inherent complexity behind the operation of ad platforms, and given the tons of personal data they use as raw material, privacy risks in the online advertising ecosystem could be studied from multiple perspectives. Naturally, most of the efforts unveiling these privacy issues concentrate on a specific entity, technology, behavior or context where user privacy is put at risk. However, such a segmented approach might underestimate the benefits of a wider vision of a problem that is systemic.

In this line, a lot of privacy protection mechanisms have been proposed from the industry and academia. The most popular ones resort to radical strategies that hinder the ad distribution process, thus seriously affecting the online advertising ecosystem (and its benefits). Others involve significantly changing the ecosystem which unfortunately may not be suitable in these times. Consequently, to encourage the adoption of privacy protection in this context, it is fundamental to pose mechanisms that aim at balancing the trade-off between user privacy and the web business model.

First, in this thesis we deal with the need to have a wide perspective of the privacy risks for users within the online advertising ecosystem and the protection approaches available. For this, we survey the online advertising infrastructure and its supporting technologies, and present a thorough overview of the underlying privacy risks and the solutions that may mitigate them. Through a systematic effort, we analyze the threats and potential privacy attackers in this scenario of online advertising. In particular, we examine the main components of the advertising infrastructure in terms of tracking capabilities, data collection, aggregation level and privacy risk, and overview the tracking and data-sharing technologies employed by these components. Then, we conduct a comprehensive survey of the most relevant privacy mechanisms, and classify and compare them on the basis of their privacy guarantees and impact on the Web.

Subsequently, we study the privacy risks derived from real-time bidding, a key enabling technology of modern online advertising. In essence, we experimentally explore how the process of user data sharing, necessary to support the auction-based system in online advertising, could be abused, at a very low cost and at the expense of user privacy. To mitigate this abuse, we propose a system to regulate the distribution user tracking data to potentially interested entities, depending on their previous behavior. This consists in minimizing the number of advertising agencies to which user data is shared while leaving unchanged the current architecture and protocols. Doing so may have an evident impact on the ad platform's revenue, thus the proposed system is designed accordingly to ensure the revenue is maximized while the abuse by advertising agencies is prevented to a large degree. Experimentally, the results of evaluation seem to suggest that this system is able to correct misbehaving entities, consequently enhancing user privacy.

Finally, we analyze the impact of online advertising and tracking from the particular perspective of Iberoamerica. We study the third-party and ad tracking triggered within local websites in this heterogenous region not previously studied. We found out that user location in this context would affect user privacy since the intensity of third-party traffic, including advertising related flows of information, varies from country to country when local web traffic is simulated, although the total number of entities behind this traffic seems stable in this context. The potential online tracking varies also when visits are simulated from European to Latin American countries and it increases when top-world sites are tested. The type of content served by websites is also a parameter affecting the level of third-party tracking: publishers associated with `news/media` or `shopping/travel` categories generate more third-party traffic and such intensity is exacerbated for top-world sites. We corroborate some of these findings and others, related to the concentration of third-party traffic in a few companies, by exploiting the transparency advertising standard `ads.txt`.

Acknowledgments

A TI, por el “buen viento”. A mis tutores por su sabiduría. A A. por el amor y el sacrificio. A mis padres por todo. A A.A. por la inspiración.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Advertising is an activity as old as commerce because, naturally, any product has to be promoted first to then be more easily sold. Such promotion entails establishing a communication channel with customers that in the past used media such as radio and TV. However, the technological progress in recent times has brought automated and intelligent mechanisms to capture the attention of potential customers worldwide through the Web.

Taken to an online context, advertising has developed its maximum potential thanks to the multiple capabilities offered. The fundamental innovation consisted in enabling advertising platforms to effectively select and deliver information to the right potential customers [2–4]. Implemented through a programmatic strategy, this approach has revolutionized the advertising business to the point that some of the biggest tech companies in the world (Google and Facebook) obtain much of their revenues from advertising. Also, to some extent, the free content created by millions of websites along the world is financed by online advertising.

Personalization is one of the key characteristics of modern online advertising. This implies distributing ads according to the specific interests of users, which is significantly more effective than broadcasting untargeted ads. Another fundamental aspect of online advertising is its efficiency in generating revenue through automated auctions, thus selling ad spaces only to the highest bidder. Evidently, real-time processes are incorporated to auction user impressions and distribute personalized

ads while a web page is displayed. Furthermore, several interfaces are provided to advertisers and other intermediary entities to enable transparency and to give them granular control over their bid strategy.

Having billions of users reachable, a very complex infrastructure has been established to implement the aforementioned services. Not only advertisers and websites are involved in this infrastructure but also more specialized entities that implement complex capabilities such as personalization or automated auctioning and that coordinate the operation among the different components of the online advertising ecosystem.

The raw material of personalized online advertising is user information. Its exploitation allows to unveil user behavior and infer potential interests. This is an important insight advertising platforms employ to direct ads to the right users. Its great effectiveness is based on the tons of data collected from users browsing the Web.

Through different interfaces, part of this information is traded from specialized data marketers. In addition, while a user impression is auctioned, his information is shared with the advertising agencies interested in order to help them guide their bid decisions. The current structure of the online advertising landscape is thus strongly dependent on releasing user data to third-parties.

Unfortunately, such user data involves sensitive attributes whose release and misuse imply serious security and privacy risks [5–7], in particular because hundreds of intermediary entities are involved and receive this information when participating in auctions. In addition to data disclosure, other characteristics of the online advertising ecosystem exacerbate privacy concerns. Mainly, the mechanisms that make personalized advertising so effective can be decidedly intrusive, to the point that users perceive said intrusion as degradation of their browsing experience [8–10]. For instance, online tracking of users along the Web, to build granular behavioral profiles, is massively performed in advertising platforms, to target potential customers, without their consent and even their knowledge. These profiles are built, among other things, on information derived from the users' browsing activity, e.g., browsing history, IP address of the user device, operating system, or plug-ins installed [4, 11, 12].

With the aim of maximizing profit, the online advertising ecosystem has facilitated the exploitation of user data, but has also encouraged the participation of advertising agencies. In this context, when the advertising platform calls for participation in the auction of a user impression, the distribution of user data that potential participants receive is basically unregulated.

This lack of regulation along with the pervasiveness of online tracking and advertising provide fertile ground for the implementation of a massive surveillance platform. Besides, despite the transparency provided to advertising entities, online advertising is supported by a completely opaque infrastructure in front of users. Namely, no information or control are granted by default for users to manage their data in this scenario.

Interestingly, beyond the release of user data, privacy risks essentially derive from the potential misuse of this data when flowing through the advertising ecosystem [13]. Such misuse might lead to characterize users as more relevant than others, depending on their behavior [14]; such a differentiation may provoke discrimination [15], a natural effect of privacy violation.

As privacy risks were not enough, the adoption of protection mechanisms in this landscape is discouraged by the complexity and opacity behind the whole structure of ad platforms. Namely, several protection approaches from the academia are unfeasible in practice since its implementation would significantly change such structure and, then, the economic model of online advertising platforms. Moreover, other solutions released by the industry, although applicable and very popular in current times, aim at directly hindering third-party interactions from the user side. This radical approach, effective and highly accepted among users, also entails a serious threat for the economic model of the online advertising ecosystem, and thus of that of the Web.

Since privacy risks in this scenario are intricate and the protection efforts so fragmented or sometimes impractical, a deep analysis of the involved entities, relationships and parameters could help to better characterize the threat scenario and thus to propose more long-term and constructive solutions for the privacy problem.

When studying this issue, we have to recognize that the lack of information regarding the internal processes of ad platforms might complicate an experimental research

on privacy threats. However, third-party interactions from the user's browser are easier to study and could reveal interesting notions of how online tracking and advertising could impact on the privacy of users, and how such impact is dependent on external influences, e.g., legislation or users' location.

Addressing these issues could help us not only unveil the serious risks behind online advertising but also build a more privacy-friendly ecosystem.

1.1 Objectives

Inspired by the aforementioned issues of the online advertising ecosystem, in this dissertation we have two main objectives. The first one is oriented to research on the privacy risks inherent to the online advertising ecosystem. The second objective encompasses designing a mechanism aimed at protecting user privacy while not significantly affecting the current revenue model. To accomplish these objectives, we first develop a general but comprehensive analysis of the privacy risks in this complex scenario. A general and systematic overview of the available protection mechanisms will also be performed. Moreover, in a more specific effort, we investigate the privacy threats derived from distributing personal information within the core of online advertising platforms. Given the automated and internal mechanism driving this process and how easily third parties get involved, serious concerns arise. To solve this particular problem, we design a mechanism to regulate the distribution of user data in online ad platforms according to the previous behavior of participant third parties. This approach will consider not only limiting the number of third parties, but also the maximization of ad platform's revenue. Finally, we experimentally analyze potential privacy risks related to online tracking and advertising from the perspective of the third-party interactions triggered from the user side. Bounded to a geographical context not explored yet, we verify whether parameters such as legislation or location affect the exposition of users to online tracking and advertising.

Bellow we provide with more details regarding these objectives.

- **Privacy risks.** We survey the online advertising infrastructure and its supporting technologies with the aim of performing a thorough overview of the

underlying privacy risks and the solutions that may mitigate them. For this, we examine the main components of the advertising ecosystem in terms of tracking capabilities, data collection, aggregation level and privacy risk, and overview the tracking and data-sharing technologies employed by these components. Based on this first approach, we conduct a comprehensive survey of the most relevant privacy mechanisms, and classify and compare them on the basis of their privacy guarantees and impact on the Web.

We also investigate, more experimentally, the privacy risks derived from the real-time auction of user impressions within ad platforms. In particular, this analysis will be focused on the unregulated distribution of information, including user data, to advertising agencies, which could lead to the creation of a massive surveillance structure on top of the online advertising ecosystem.

Still in line with revealing privacy risks, we concentrate on the user side and the third-party interactions triggered from there when browsing the Web. By studying such communications, we measure the influence of online tracking and advertising, particularly on Iberoamerican countries.

- **Privacy protection.** We design and evaluate a system to mitigate the impact on privacy provoked by the unregulated distribution of information to third parties within ad platforms. We propose a mechanism to restrict the participation of such third parties when ad platforms share tracking data. A strategy will be designed to preserve the revenue while potential abuse is detected.

1.2 Summary of Contributions

Below, we list the main contributions of this thesis.

- We systematically survey the current state-of-the-art of academic and industry solutions that aim at protecting Web users from various privacy threats posed by the online advertising industry. To this end, we characterize the capabilities of the components involved in the ad-delivery process, in terms of type and scope of data collection, aggregation level, and, accordingly, privacy threat.

- Motivated by the previous analysis of privacy risks, we develop a comprehensive overview of the protection mechanisms that may cope with such threats. These mechanisms are examined, among other aspects, on the basis of the location of the mechanism employed, the scope of its application and its protection strategy. We concentrate on those privacy mechanisms that operate on the user side, since the opacity of online ad platforms difficult further research inside. Our review of privacy mechanisms establishes a correspondence between the privacy risks identified and the proposals, both from academia and industry, that may address them.

- We experimentally evidence the potential misuse of real-time bidding, a key technology supporting the online advertising ecosystem, by which user data is distributed among third parties as background information in the process of auctioning a user impression. We quantify the extent to which an advertising agency may collect user tracking data without even paying for it.

- In order to cope with the aforementioned issue, we conceive a system that aims at regulating the distribution of user data to third parties during the auctions for ad-impressions, i.e., to whom send the requests for each ad-space bidding. Limiting the number of third parties receiving user profiles naturally offer better privacy protection, especially since potential dishonest entities will hardly receive user sensitive information under such context. We formulate the problem of choosing a distribution strategy as a multi-objective optimization problem that takes into account both aspects, i.e., the number of entities receiving information and ad platform's profits.

- Lastly, we investigate the privacy issues in online advertising from the perspective of the interactions triggered from the user side to ad related third-parties. We focus on a novel scenario that involves online tracking and advertising spawned by local websites in Iberoamerican countries. In particular, we aim at finding out how user location and content type served by publishers impact on potential third-party tracking and privacy risks.

1.3 Related Publications

Journal publications:

1. J. Estrada-Jiménez, J. Parra-Arnau, A. Rodríguez-Hoyos, and J. Forné, "On the regulation of personal data distribution in online advertising platforms," *Engineering Applications of Artificial Intelligence*, vol. 82, pp. 13–29, June 2019. ISSN: 0140-3664. Impact factor 2019: 4.201 [16].

2. J. Estrada-Jiménez, J. Parra-Arnau, A. Rodríguez-Hoyos, and J. Forné, "Online advertising: Analysis of privacy threats and protection approaches," *Elsevier Computer Communications*, vol. 100, pp. 32–51, March 2017. ISSN: 0140-3664. Impact factor 2017: 2.613 [17].

3. J. Estrada-Jiménez, J. Parra-Arnau, A. Rodríguez-Hoyos, and J. Forné, "Measuring Online Tracking and Advertising in Iberoamerica," to be submitted to *IEEE Access*. ISSN: 2169-3536. Impact factor 2019: 3.745.

4. A. Rodríguez-Hoyos, D. Rebollo-Monedero, J. Estrada-Jiménez, J. Forné, and L. Urquiza-Aguiar, "Preserving Empirical Data Utility in k-Anonymous Microaggregation via Linear Discriminant Analysis," accepted to be published in *Elsevier Engineering Applications of Artificial Intelligence*, May 2020. ISSN: 0952-1976. Impact factor 2019: 4.201 [19].

5. E. Pallarès, D. Rebollo-Monedero, A. Rodríguez-Hoyos, J. Estrada-Jiménez, A. Mohamad Mezher, and J. Forné, "Mathematically optimized, recursive prepartitioning strategies for k-anonymous microaggregation of large-scale datasets," *Elsevier Expert Systems with Applications*, vol. 144, pp. 1–17, April 2020. ISSN: 0957-4174. Impact factor 2019: 5.452 [20].

6. A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, J. Parra- Arnau, Ahmad Mohamad Mezher, and J. Forné, "The fast MDAV (F-MDAV) algorithm: An algorithm for k-anonymous microaggregation in big data," *Elsevier Engineering Applications of Artificial Intelligence*, vol. 90, no. 103531, April 2020. ISSN: Impact factor 2019: 4.201 [21].

7. A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, "Does k-anonymous microaggregation affect machine learned macrotrends?," *IEEE Access*, vol. 6, pp. 28 258–28 277, May 2018. ISSN: 2169-3536. Impact factor 2018: 4.098 [23].

Conference publications:

1. J. Estrada-Jiménez, J. Parra-Arnau, A. Rodríguez-Hoyos, and J. Forné, "Measuring Online Tracking and Privacy Risks on Ecuadorian Websites," *IEEE Fourth Ecuador Technical Chapters Meeting (ETCM)*, Guayaquil, Ecuador, November 2019 [18].

2. A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, J. Forné, R. Trapero, A. Álvarez, and R. Rodríguez, "Anonymizing cybersecurity data in critical infrastructures: The CIPSEC approach," in *Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, Valencia, Spain, May 2019 [22].

3. A. Rodríguez-Hoyos, J. Estrada-Jiménez, L. Urquiza-Aguiar, J. Parra-Arnau, and J. Forné, "Digital hyper-transparency: leading e-government against privacy," in *Proceedings of the 2018 International Conference on eDemocracy & eGovernment (ICEDEG)*, Ambato, Ecuador, June 2018. [24]

1.4 Outline of this Thesis

We detail below the structure of this dissertation which follows the objectives defined in Sec. 1.1.

Chapter 2 introduces the main concepts with regard to the online advertising ecosystem. This includes a brief history of its evolution over time, a succinct description of the main players, and the technologies supporting ad displaying. Privacy issues derived from online advertising and related protection approaches are also introduced in this chapter.

Chapter 3 presents a systematic survey of the online advertising infrastructure, including a deep analysis of the solutions available to mitigate such risks. The main components of this ecosystem, its capabilities and the technologies supporting inherent processes are studied in this chapter. Such parameters are then matched with potential privacy abuses. Also, an analysis of several privacy protection approaches is presented along with a categorization of these tools according to their specific scope of application.

Chapter 4 proposes examining a particular privacy risk arising, within advertising platforms, from the uncontrolled distribution of personal information to third parties. Afterwards, a cost-effective mitigation mechanism is presented.

Chapter 5 is devoted to measure online tracking and advertising in the context of Iberoamerican countries. The impact of privacy risks depending on the location of users and websites are also discussed in this chapter.

Finally, conclusions are drawn in Chapter 6.

Chapter 2

Background and Related Work

2.1 Introduction

Selecting and directing information are crucial in every aspect of our modern lives, including areas as diverse as health, leisure and research. In the past, these processes were largely manual, but due to the exponential improvements in computation and sophistication of software, they are becoming increasingly automated.

The industry of online advertising, lavishly illustrated by Google DoubleClick and real-time bidding (RTB), is an example of the ever-growing automation of these processes, and another crucial aspect of our society — to a large extent, the success of most competitive economic activities is dependent on advertising, particularly on the ability to effectively select and direct information to the right potential customers.

Undoubtedly, the advent of the Internet and the Web has created a myriad of new opportunities for advertisers to target billions of people almost effortlessly. However, online advertising is not only ubiquitous. In the early days of the Web, ads were served directly by the publisher (i.e., the page's owner) following a one-size-fits-all approach. But due to the ease with which Web users can be tracked across their page visits, online advertising has also become increasingly personalized. An example of the sophistication of ad personalization is RTB, which enables advertisers to direct

Figure 2.1: Word cloud of terms related to online advertising, tracking, user profiling, and privacy solutions in this scenario. We discuss all these terms in this chapter. The font size of each of them is proportional to the frequency of occurrence in Google search.

ads to the right user and at the right time, by competing in real-time auctions for the impression of their ads [25].

Evidently, personalized advertising is the most effective, and hence the most profitable, form of advertising. Those ads relying on a user's browsing interests ensure conversion rates[(a)] that double those of untargeted ads [26]. On the other hand, from the publishers' perspective, online advertising is the pillar that sustains the Internet's "free" content and services.

Nevertheless, advertisers and publishers are not the only entities taking part in this business. In fact, there exists an entire infrastructure at the service of both of them, supported by companies like Google, Facebook and Twitter. Enabled by these and hundreds of other ad companies, targeting mechanisms take charge of selecting and directing ads to billions of users everyday, depending on a number of factors such as the page they are visiting; their browsing history; their IP address or parts of it; their operating system; the plug-ins installed and other information related to their Web browser [4, 11, 12]; and obviously the objectives and budgets of all advertisers for displaying their ads.

User information is therefore an asset fundamental to the efficient and effective delivery of advertising, which is not only handed over to the highest bidder, but to many other third parties that are involved in the ad-delivery process. Unfortunately,

[(a)] In online marketing terminology, conversion usually means the act of converting Web site visitors into paying customers.

evident security risks exist for users when personal, sensitive data about their habits are traded in the name of personalized advertising by an infrastructure that operates in the shadows with virtually no oversight [27]. These security risks can be explained in terms of privacy hazards, social sorting, discrimination, malware distribution, fraud and others [5] [6] [7].

Regarding privacy, serious concerns have been raised by the intrusiveness of practices and the increasing invasiveness of digital advertising. According to recent surveys, two out of three Internet users are worried about the fact that their online behavior be scrutinized without their knowledge and consent. Numerous studies in this same line reflect the growing level of ubiquity and abuse of advertising, which is perceived by users as a significant degradation of their browsing experience [8] [9] [10].

In an attempt to mitigate these privacy and security risks, several approaches have been proposed by a heterogeneous group of actors. Research proposals have concentrated on sophisticated mechanisms to anonymize or block the information leaked to third-parties while trying to remain compatible with the current ecosystem. On the other hand, commercial solutions have primarily focused on blocking tracking mechanisms at the cost of seriously damaging the Internet business model.

2.2 Evolution of the Online Advertising Landscape

Advertising is commonly linked to commercial activities that involve branding strategies intended to draw the attention of potential customers. The objective of drawing attention is persuading users to buy a product or, generally, spawning brand image. Historically, however, the way potential customers have been contacted by advertisers to apply such strategies has ended up bothering the ones they aimed at attracting [28].

The main problem of classical online advertising has been commonly the very limited media infrastructure by which ads have been distributed to customers. Without enough resources to target users (e.g., TV viewers or newspaper readers), advertisers used to massively flood the available media with ads which very few people were interested in [29]. The flooded message usually “touched” some customers but the strategy

was definitely inefficient. Currently, marketing announcements are still sent to an audience that has a huge aggregate size but which is also ultra-fragmented [30] [31]. This is due to the broad range of available media channels (TV channels, websites, etc.) and the volatility of the attention users put on such channels [32].

Despite its shortcomings, online advertising has been a profitable business and proved to be effective in terms of ROI [(b)], interaction and tracing of potential customers, and reaching an audience [33]. The truth is also that, in the past, audiences were not as fragmented, and the online ecosystem was not as congested as it is currently. As a result, there were more chances for such traditional advertising strategies to be successful.

With the rise of the Internet, the advertising industry has evolved significantly, especially in terms of its capability of reaching potential customers on an individual basis. Modern online advertising takes advantage of recommendation and personalized information systems to tailor advertising campaigns to the interests of Web users [34]. Thus, thanks to technologies like RTB, the core of the advertising business is able to show ads to the right person and at the right time, which implies greater effectiveness [7, 35, 36]. Additionally, current online advertising provides more accountability and transparency since ad companies are encouraged to agree on prices that directly match the effort undertaken by the seller with the benefits received by the buyer. Consequently, in economic terms, advertising services are traded based on the force of demand and supply [4].

Although online media have transformed the way advertising is conceived, it was not always so. The online environment was originally overwhelmed by confusion where the impact and fulfillment of advertising campaigns were hardly determined objectively [7, 11]. For instance, advertisers had to acquire inventory of spaces available to publish ads without really knowing if such spaces were shown to people interested in the promoted products. Moreover, the lack of resources of the emerging advertising technologies of that time prevented online actors from optimizing the ad-delivery process.

[(b)] ROI or return on investment is an indicator used to measure the efficiency of an investment.

At present, the online advertising landscape is triggered by advertisers, who create the demand, and publishers, who generate the supply. Websites have become the publishers by excellence since the content they offer attracts people whose interests can be revealed from intrinsic interactions with the Web. Moreover, modern methods of online advertising management have incorporated intermediate entities that help advertisers and publishers navigate the web topology in order to connect them together [7]. Such intermediaries, as explained below, are responsible for providing interactive and automatic ad serving that is able to accurately target the intended audience. The targeting strategy implemented by these intermediary entities has directly influenced the ad-personalization accuracy, but also the level of transparency of the process whereby ads are delivered.

Lastly, it is worth stressing that the money produced by online advertising is currently sustaining most of the "free" content on the Web [37]. The money paid by advertisers becomes revenues that are distributed among the different actors of the ecosystem, including the publisher [11].

2.3 Online Advertising Players

The modern online advertising infrastructure has become certainly complex and dynamic and, although more players can be identified, three components deploy the main roles in this industry. As illustrated in Fig. 2.2, these components are advertisers, publishers and ad platforms, and their ultimate goal is to display the right ad to the right user [4] [38]. The former two components represent respectively the demand and supply sides of the economic model that governs an online advertising service [7]. The interactions between such players are commonly enabled by an intermediate infrastructure called ad platform. Finally, users, whose data and requests are the basis of the decisions made for online advertising services, are not directly considered as part of this infrastructure since they do not receive the revenues of such billion-dollar business.

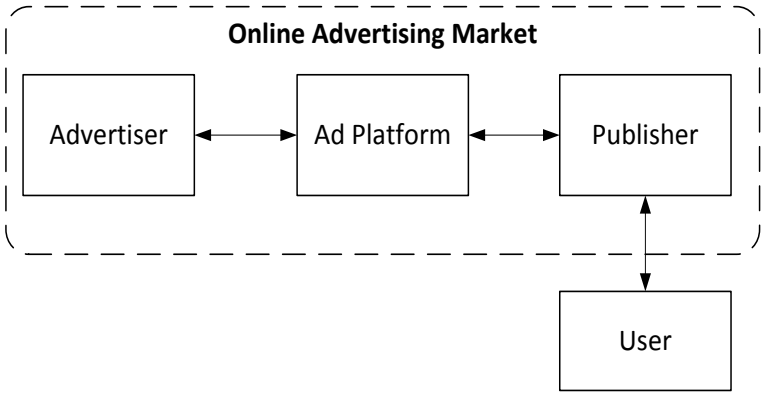


Figure 2.2: Main components of the online advertising ecosystem.

Advertisers are entities that are interested in promoting a brand or product by showing related ads to potential customers. They are willing to pay for displaying their ads [4] [38], and therefore they are the entities that generate the demand of advertising services. Online advertisers are basically aimed at displaying ads on some spaces of the websites (publishers) users visit. Direct agreements may be signed among advertisers and publishers to regulate the online ad service, but these actors commonly get engaged through intermediate platforms, as shown in Fig. 2.2. Obviously, the use of intermediary entities makes this process more efficient. Thanks to these entities, advertisers may target ads to the intended audience of their marketing campaigns. Also, through modern online advertising mechanisms like RTB, they may participate directly in this targeting process. These capabilities are crucial for advertisers to face the fragmentation of online audiences.

A **publisher** is an entity, such as CNN or The New York Times, which provides online content (e.g., newspapers, search engines, blogs, etc.), usually through web pages. Since such content draws the attention of users, advertisers pay publishers to be assigned a space in a website, where they can show ads to a given audience. Commonly, publishers supply advertisers with an inventory of spaces (on their websites) to be filled with marketing messages. Such inventory can be sold by contract or in real time. As depicted in Fig. 2.3, a publisher is the entity through which a user comes into contact with the online advertising ecosystem.

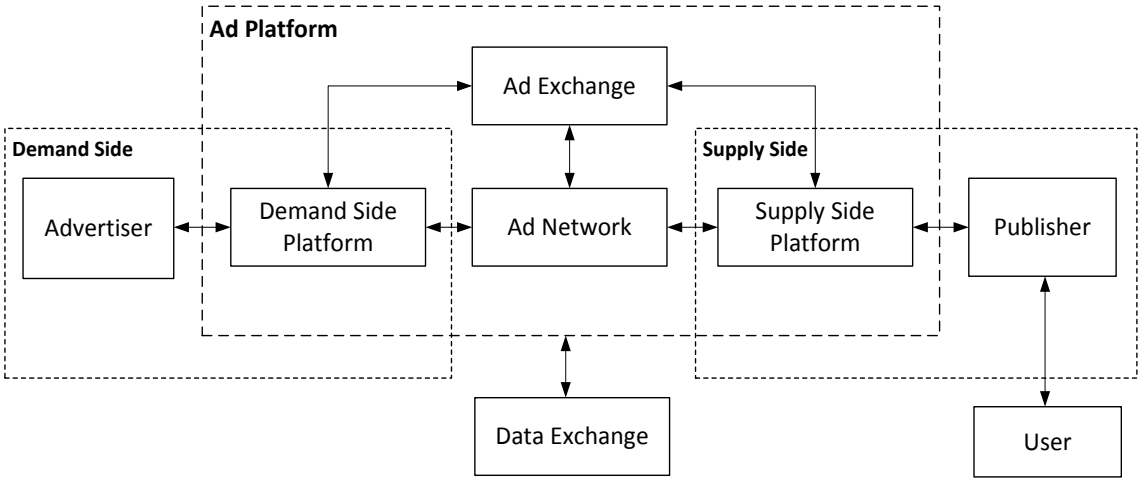


Figure 2.3: Disaggregated ad platform scheme and interactions between players.

Ad platforms are groups of entities that connect advertisers with publishers through their demand and supply-side interfaces. In particular, as can be seen in Fig. 2.3, ad platforms constitute the marketplace where the demand and the supply of online advertising services are matched [4]. In order to effectively reach the currently fragmented online audiences (i.e., a multitude of websites and a pretty scattered attention of users), ad platforms arose to help advertisers and publishers increase the selectivity and efficiency of ad space allocation. Therefore, ad platforms may be considered as the centerpiece of the modern Internet advertising business as they facilitate the matching between the advertising material and users' interests. The accuracy of said matching clearly depends on the ad platforms' ability to track and profile users based on the information that can be mined from their online activity. The ad-targeting process has in recent years become increasingly sophisticated, which has inevitably led to the emergence of numerous agents with very specialized roles. The upshot of this more populated ecosystem (see Fig. 2.3) is a more automatic, transparent and flexible ad-delivery process. Throughout this work, we refer to ad platforms as *all* the intermediary entities that connect advertisers to publishers.

Originally, ad platforms used to aggregate only the inventory provided by publishers. The aim was to help advertisers get scale and impact (in terms of amount) when distributing their ads; however, scale was not enough. Later, modern ad platforms brought a more transparent infrastructure where advertisers became capable

of selecting the users to which they wanted to show ads. To this end, ad platforms integrated certain mechanisms to make the ad-targeting process more accurate, transparent and flexible. Such mechanisms are now implemented by different entities that are part of ad platforms. These entities provide complementary services including aggregation of demand and supply, and optimization of the ad-serving process itself. Some of these entities are *ad networks*, *ad exchanges*, and *demand and supply-side platforms* [7]. Ad networks and ad exchanges are the predecessors of ad platforms. Ad networks began aggregating inventory for advertisers, and ad exchanges evolved to include more dynamic mechanisms to serve ads through automated auctions [39].

Ad networks emerged to help advertisers select and buy ad spaces across the congested and fragmented ad-serving infrastructure. With this aim, such networks used to resell the aggregated ad inventory acquired from publishers to advertisers and related agencies [39]. For those publishers that directly sold their inventory to big advertisers, ad networks became an interesting entity through which to sell their remnant inventory for a good price [11]. Other smaller ad networks were able to give advertisers access to more selective audiences by aggregating more specific inventory from small publishers. Examples of ad networks include Google AdSense, Media.net and PulsePoint.

Ad exchanges are ad platforms that currently sell their aggregated inventory of ad spaces by means of auctions. They keep consolidating ad spaces from publishers but offer advertisers and publishers more effective and transparent mechanisms to serve ads [4, 40]. First, ad exchanges place ads based on automated auctions where advertisers "decide" how much to pay for an ad space. The winning bidder is the advertiser that ends up displaying the ad. Secondly, during the auction, ad exchanges share with advertisers contextual information about the user who generates the impression they bid for. Such information helps advertisers decide whether to bid for an ad space and how much to bid for it. The auction is held just after a user requests content from a website partnering with the ad exchange. The whole process may take a few tenths of a second. Theoretically, this yields greater efficiency since the ad-delivery process is distributed among the different components of the ad platform [11]. Part of the aggregation strategy of ad exchanges consists in combining

multiple ad networks together. This way, advertisers and publishers are relieved from dealing with so many intermediaries.

Demand-side platforms (DSPs) are entities that work for advertisers, i.e., for the actors generating the demand of ad services. DSPs work on behalf of advertisers, in front of the ad exchange, and help advertisers choose audiences and adequate media to display their ads. By aggregating demand, DSPs are capable of boosting selectiveness and effectiveness for advertisers [4, 11].

Supply-side platforms (SSPs) are entities that work on behalf of publishers, the actors that supply ad spaces to advertisers. SSPs offer publishers an optimized strategy to manage their advertising inventory. Since the task of targeting an ad to a given user involves advanced capabilities and resources, publishers delegate this task to SSPs, with the hope of getting increased demand and profits, despite the congested online ecosystem.

Data aggregators are entities that collect information about Internet users with the aim of profiling their purchasing interests. Data aggregators' services aim at tailoring ad marketing strategies to the users' preferences they have learned by means of massive data mining. From data aggregators, another entity called *data exchange* arises. Data exchanges provide demand and supply-side platforms as well as ad exchanges with user data to help them make their targeting decisions.

General Operation of Online Advertising

Having shown the main components of the online advertising ecosystem, now we proceed to briefly describe how ads are delivered on the Web.

Currently, ad serving aims at providing automated processes and transparent interactions to advertising entities. However, there are many interactions involved that make the ad-serving process really complex and completely opaque to the user. In general, when a user visits a website, personalized advertisements are displayed together with the content of the site, as if they were part of the same structure. According to the user's perception, ads seem to be served by the same web server.

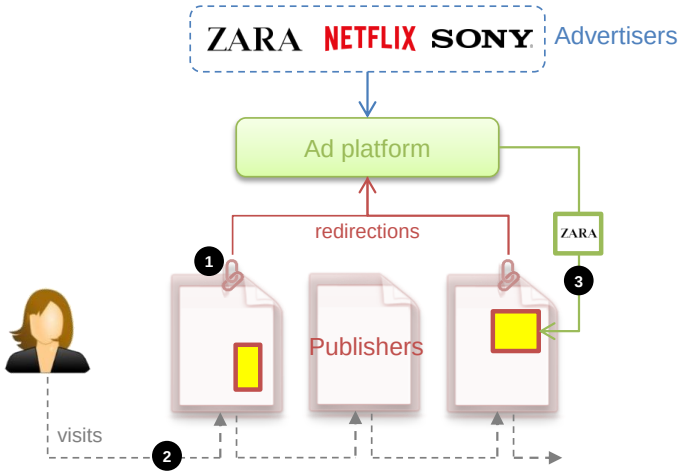


Figure 2.4: Current online advertising architecture composed by publishers, ad platforms and advertisers.

Although the user participation in the ad-serving process is merely passive, the entire process is triggered by a user's request to download Web content.

This way, when a user's browser sends an HTTP request to a website that is associated with an ad exchange, the website sends back the content the user is requesting. Such content is interpreted by the browser and then displayed to the user. Along with the content, additional code, in the form of ad tags, is sent to the browser and executed automatically. The execution of this code triggers a connection from the browser to the ad exchange in question, which asks for advertisements to fill the ad spaces on the visited page. When the ad exchange receives the ad call, the process of selecting the right ad for the best price is performed by some of the intermediary entities described above. Mechanisms such as RTB and *cookie matching* (CM) are used to ensure the greatest impact on users (which benefits advertisers) together with the highest profits for the ad-serving platform (which includes publishers). Figure 2.4 shows the current architecture of online advertising composed mainly by publishers, ad platforms and advertisers, and illustrates the process whereby third-party ads are displayed to users.

The ad-delivering process requires that publishers include a link to the ad platform they want to partner with (1); for the sake of simplicity, we consider here a single ad platform. When a user visits pages partnering with this ad platform, the browser is

instructed to load the URLs provided by the ad platform. Through the use of third-party cookies and other tracking mechanisms, the ad platform is able to track all these visits and build a browsing profile (2). Based on this profile, the user's location and other parameters, the ad platform uses its targeting algorithm to decide which ad to present on the publisher's page

2.4 Supporting Technologies

The ad-serving process has significantly evolved from the days when advertisers selected the media to deploy ads long before a user visited a website. Currently, advertisers may decide, in real time, which ad to display. As described in Sec. 2.3, ad platforms take in the order of milliseconds to target an ad to a user based on their preferences and the campaign requirements specified by the advertiser in question. Two main processes are involved. On the one hand, a behavioral profiling task is conducted against a visiting user; this is done on the basis of any information collected about them [35]. On the other hand, automated auctions are used to distribute ads in favor of advertisers, in accordance with their willingness to bid for a particular profiled user.

Although behavioral profiling and real-time auctioning are key processes for modern online advertising, they would not be feasible without *online tracking*. Online tracking enables the collection of the user information that is afterwards employed as input for user profiling and bidding decision making.

Mechanisms such as CM and RTB have been developed to support the modern online advertising platforms, by facilitating ad serving personalization and enabling a more efficient and profitable ad distribution system. In the coming subsections, we overview these two mechanisms and online tracking.

Cookie Matching

In order to decide whether and how much to bid for users' impressions, online advertisers require as much information as possible about such users. To come to that

decision, the first task of ad platforms is to individuate users so that different attributes can be associated with a (almost) single virtual identity. CM is a mechanism that assists an online advertising platform, and in general a web tracker, in "recognizing" users across the Web. As we explain later on, said assistance is key to the bidding processes [41].

CM is based on cookies, which are randomly generated strings of text that web servers send to users' browsers. Cookies are employed to recognize users in subsequent visits. By "identifying" their users, servers are capable of offering personalized services. The same strategy is applied by an ad exchange when serving ads to users, in order to recognize them on a later auction. When a new auction is to be held, an ad exchange sends (ad call) the identifier it keeps about the user to the prospective bidders (advertisers). Such an identifier (cookie) allows advertisers (or their corresponding DSPs) to find any other cookie left on the user's browser in previous auctions. Moreover, an advertiser by itself might have placed cookies on the user's browser from a process unrelated to auctions [11]. Cookies coupled with auction processes may enable advertisers (and other entities) to build profiles of users with information about their browsing history and buying habits.

The process of CM, also called cookie syncing, allows an advertiser and an ad exchange to match the identifiers (cookies) they have about a single user, so that they can share information about them. As stated above, such information enables advertisers to make a more informed decision on whether and how much to bid for an ad impression. A detailed description of how CM works in Google's ad exchange DoubleClick can be found in [41].

Real-Time Bidding

Bidding, in general, has represented a breakthrough for the online advertising business. Bidding initially arose for paid-search advertising [42], with the aim of giving transparency to the process of ranking advertisers on search engine results pages. After spamming had affected the quality of search results provided by search engine marketing, and after having realized that such a system prevented smaller companies

from participating in the emerging online advertising system, auctions appeared as a mechanism to "democratize" the access to the ad-serving ecosystem [11].

RTB, also called programmatic buying, is an auction-based technology for online advertising. RTB mimics a stock exchange to enable automatic buying and selling of ads [25]. This automatization allows RTB to perform a per-impression bidding just in the moment such an impression is generated. Classic bidding used to take place way before the user accessed the web page where an ad was displayed. Modern bidding, however, is perceived as a real-time process since ad serving is conducted in a fraction of second [43].

RTB enables advertisers to bid for the chance to display an ad on a web page loaded by a user's browser. After such a process, a publisher shows the ad of the advertiser that won the bid. When a user spawns a request from their browser to a website engaged with an ad exchange, a corresponding ad call is generated to the ad exchange. Upon receiving the ad call (asking for advertising), the ad exchange sends a bid request to the advertisers that might be interested in sending ads to a user. Along with the bid request, ad exchanges send valuable information about the user whose impression is being auctioned [14]. Cookies are extensively used by ad exchanges and advertisers to collect and share such information, and thus improve the accuracy of the ad-targeting process [44]. In fact, the very detailed contextual information provided through cookie-related technology helps advertisers and DSPs to make the decision of whether and how much to bid for an impression. After bids are made, a winner is determined during a real-time auction. In a last step, the ad exchange notifies the winner advertiser and its ad is served on the website through the user's browser. This last step may entail a content-delivery network.

Online Tracking

As users browse the Web, they are monitored by the entities they interact with. This is possible due to user information leaked in the requests sent to such entities.

These entities involve "first parties," which are the websites directly visited by users, and "third parties," which receive user requests automatically (and inadvertently) triggered from the user's browser. While first parties may be seen as the

intended receivers of user requests, third parties are usually hidden trackers, such as ad networks, contacted by users as a result of code embedded on most web pages.

Since third parties may be coupled with several web sites, they could receive interactions of the same user in multiple contexts, e.g., when visiting different web sites. Besides, the information of the first-party site is released to the third party in the "referer" header, but other, more sensitive, information could also be released. Accordingly, third parties could build user browsing histories if supported by other tracking technologies such as cookies or CM to individuate users.

In a nutshell, online tracking consists in "following" users along the Web, i.e., basically knowing when, from where, and what users visit. This is done by triggering automatic connections from users to third parties when they visit a publisher. Needless to say, such connections carry several information items about users, which allow external entities to become potential trackers.

2.5 Privacy Issues

Several privacy concerns arise from the complex infrastructure developed to support online advertising. Undoubtedly, the massive collection of user data and its potential misuse by third parties is one of the most critical.

However, massive collection is motivated by the great utility that can be extracted currently from big data. In the case of online advertising, data exploitation for distributing personalized ads supports a billionaire business where even small companies can participate.

The advertising ecosystem is then crowded with thousands of third-party entities avid of exploiting personal data for profit. Such amount of potential privacy attackers is hard to be controlled, even more if the asset to be protected (information) is the raw material of a business that is said to be supporting the current free-content distribution model on the Internet.

This whole ecosystem grows exponentially together with the Web, mainly as a result of an increased online access rate of the world population. That growth, as well as the need to meet stricter real-time application requirements, has given rise

to a very complex infrastructure. Such complexity has left little interest in facing privacy risks.

Finally, despite the recent regulatory efforts for personal data protection, the strong economic interests behind online advertising keep hindering the effective implementation of privacy protection initiatives. For instance, a series of "dark patterns" are being adopted by publishers and third parties to circumvent their transparency and consent management obligations.

Next chapter addresses more deeply these privacy issues.

2.5.1 Privacy Protection Approaches

In general, the concerns regarding privacy arise from the inappropriate collection, use and sharing of user data [13]. In the context of online advertising, said misuse of user data is potentially present in different moments in time. First, powerful tracking mechanisms are employed by high-level advertising players to "follow" users through the Web [40,45,46]. These tracking mechanisms include cookie matching and fingerprinting whose implications will be analyzed in the next chapter. When users navigate a website serving ads, third-party interactions from the browser disclose user data to said players, which aggregate and store this information (collection). Then, they process this user data (use) and further distribute it (sharing) to enable personalization for users and to guide the targeting strategy of advertisers.

Within this framework, *external* control over the flow of data could only be enforced before the collection step, i.e., when the web browser leaks data in third-party interactions. Further adjustments require changing the online advertising structure. This is why most of the functional solutions to protect privacy in this domain build on managing (essentially detecting and blocking) third-party connections from the user side. These are local approaches, commonly implemented as web browser extensions, that provide users with tracking blocking capabilities. The most popular ad blocker is AdBlock Plus [47], but other similar tools exist that also provide transparency and user personalization [48–50]. In this line, other initiatives propose blocking strategies implemented in brokers [51] [52] [53] that act as local proxies to filter the interactions

performed between a group of local users and advertising entities on the Web. Historically, these approaches have detected third-party tracking through static blocking lists that have become extremely long and hard to manage [54], but recent proposals have improved such detection by using machine-learning techniques [55].

However, ad blockers and anti-trackers suffer from controversial shortcomings. First, its extended use is seriously threatening the business model of the whole Internet. Also, though radical and apparently infallible, ad blocking would have been circumvented by tracking companies by exploiting web sockets [56]. Namely, ad blockers might not be as effective as expected.

Looking for more advertising-friendly solutions to preserve privacy, multiple initiatives have emerged from the academic world. Those mostly suggest integrating the active participation of users so they can decide how to manage their data. Some of these works [51] propose incorporating trusted third-parties to intermediate the communication between users and advertising players to encrypt or obfuscate user data. Several other approaches present advertising architectures where the exploitation and sharing of data is moved to the user premises, i.e., to the user's browser or a local application [52, 57, 58]. This enables users to control how their data is processed and how and when it is shared to third-parties. Two recent research works [59, 60] present protocols to exploit personal data while auctioning user impressions without revealing any personal preferences (in clear text) to advertising parties. As some of the previous approaches, these protocols require that user information be processed locally in the browser, and that a trusted third-party assist in performing operations over encrypted user data.

Other more revolutionary proposals even suggest adapting the advertising model to allow users to be rewarded for ceding their data [61, 62] or to enable advertising players to charge users for not tracking them [63]. Sadly, all this related academic research requires modifying the current online advertising model, either in the way user data is exploited or in the mechanism to obtain (economic) value from it. These are important changes that would significantly impact the utility of the user information received by ad platforms, thus negatively affecting their huge revenues. As a

consequence, there might not be incentives for the advertising market to adopt them in the short-medium term.

With blocking solutions that are critically tampering with the economy of the Web [64] and academic approaches that are not feasible in the short term, it seems that we need to look beyond to get real privacy. Reaching effective strategies implies starting to disrupt the core of the ecosystem in order to address the moments when user data is processed and shared. Some steps in that line are already taking place thanks to strict privacy regulation recently promulgated [65] in Europe that is motivating companies to cooperate in favor of the privacy of users. Interestingly, the mere application of transparency initiatives has already allowed to unveil further privacy risks within advertising platforms [66, 67].

Chapter 3

Privacy Threats and Protection Approaches in Online Advertising

3.1 Introduction

In Sec. 2.5, we commented that the pervasive dissemination of online advertising on the Internet and the prevailing need of ad platforms and other intermediary entities to collect a wealth of data about Web users prompt serious concerns regarding user privacy [68, 69]. In fact, much of the concern regarding privacy and thus regarding privacy threats in online advertising are derived from the risks of misuse of this huge amount of user data, which is held by advertising platforms. Said misuse of user information might include common privacy issues such as data leakage, unauthorized collection of data, and sharing with a third-party. Interestingly, as surveyed in Sec. 2.3, the structure of ad platforms and the abilities of their players reflect behaviors strictly coincidental with such privacy issues.

In accordance with the above reflection, in this chapter, we first identify the privacy threats specifically inherent to, or arising as a result of, online advertising, based on a characterization of the main players as potential attackers, and of the effects of their capabilities as primary threats. Note that, although the concept of privacy is intimately related to that of information security, the former is addressed here as a particular field of the latter, whose focus is on protecting user data from

being revealed, without consent, to potential attackers. Thus, the scenarios in which the user information leaks could be classified as risky.

The privacy risks posed by the tracking and profiling practices of the online advertising industry have motivated a variety of privacy-protecting approaches from academia. These research initiatives mostly rely on mechanisms that may support or complement the current economic model of the Web, while others suggest moderate blocking of third-party tracking[(a)] to protect user privacy. Other plug-and-play proposals are also available to users and are supported commercially. In essence, such approaches provide users with transparency and control functionalities over their browsing data, yet putting at risk the Web economic model, currently built on the revenues of online advertising, through radical blocking mechanisms.

In this chapter, we address the main parameters that characterize the current privacy protection approaches in online advertising, in particular, their location, scope of application and strategy. Afterwards, we analyze the most relevant research work and industry proposals which tackle the problem of privacy protection in online advertising.

This analysis exclude the specific context involving mobile devices, albeit much of the following reasoning might still be true for both desktop-based and mobile browsing. Certainly, advertising in mobile communication environments, deserve a separate study, given the complexity of their infrastructures and the growing use of smartphones connected to Internet.

In the next paragraphs we provide more detail regarding the objectives of this chapter.

- This chapter presents a "big picture" of the state-of-the-art of academic and industry solutions designed to protect web users from privacy threats posed by the online advertising ecosystem. This is done by characterizing the main actors of this infrastructure, the interactions among them, and their supporting technologies. Said introduction provides with the necessary depth to understand the intricate dynamics of the current advertising ecosystem, and the privacy risks users are exposed to.

[(a)] The vast majority of ads today are served by third-party entities [2, 70].

- To shed some light over these risks, we conduct a thorough analysis of the capabilities of the components involved in the ad-delivery process, in terms of type and scope of data collection, aggregation level, and, accordingly, privacy threat. This allows the definition of an adversary model and a systematic classification of the elements of the online advertising architecture.

- In addition we perform a comprehensive overview of the protection mechanisms that may cope with such threats. These mechanisms are examined, among other aspects, on the basis of the location of the mechanism employed, the scope of its application and its protection strategy. First we concentrate on privacy mechanisms that operate on the user side, since the opacity of online ad platforms has not allowed further research inside. This review establishes a correspondence between the privacy risks identified in the first part of this chapter and the proposals, both from academia and industry, that may address them.

By systematizing the analysis of privacy risks and protection mechanisms, this chapter aims at providing privacy designers and researchers with a far-reaching picture of the current state of affairs in online advertising.

The work presented in this chapter was published in [17].

Chapter outline

The rest of this chapter is organized as follows. Section 3.2 examines the privacy risks inherent to the scenario of the online advertising ecosystem. Section 3.3 conducts a thorough analysis of the most relevant mechanisms to mitigate such risks. In Sec. 3.4, we discuss the various threats identified and the mechanisms that may address them. Finally, conclusions are drawn in Sec. 3.5.

3.2 Privacy Threats in Online Advertising

3.2.1 Attacker Model

Privacy criteria are commonly defined in terms of the amount and quality of information that potential attackers might be able to collect about users. Further, characterizing such potential attackers is of special relevance since user privacy is generally measured with respect to the adversary's capabilities as in [71].

Should we consider any entity with access to user data as a privacy attacker, the modern online ecosystem is nowadays plagued by potential adversaries. In the context we address, such adversaries are the multiple intermediate entities developed as part of the online advertising architecture. Although most of these prospective attackers are not directly involved in the raw web traffic spawned by a user, a variety of contextual user information is leaked to ad-serving entities [72] [57]. In general, the information typically collected about a user includes their clickstream, browsing history, shopping habits, preference ratings, entertainment preferences, location, gender, age, and agent string [46].

The online applications and devices (such as browsers and computers) that are daily employed by users lend themselves to the generation of a sort of digital signature that can be subject to fingerprinting. This signature is built with a chain of pieces of information (software installed, plug-ins, and version of applications) that almost uniquely identify a user on the Web. No matter if a user deletes their cookies, she can be tracked online through such a string of data, commonly called an agent string [46].

Even though these items of information might not seem relevant to the identity of a user, several studies have shown that data on some of these "tags" might be sufficient to unambiguously identify a user within a country [73] [74].

Potential attackers in the online advertising ecosystem could be classified as *first* and *third parties*, according to the interaction level of each entity with the user. A first party is directly (consciously) contacted by a user. Nevertheless, third parties are contacted through requests that are not explicitly triggered by users. In this context, publishers may be regarded as the only first-party entities, since the interaction with

them is directly made by users; the rest of the components of the advertising architecture depicted in Fig. 2.3 may be considered as "third-party adversaries". Naturally, the scope of all these potential privacy attackers will vary from local to global according to the amount of users whose information is traded through every component. Of course, such hierarchical scope will determine the aggregation ability and, therefore, the level of privacy risk posed by each of these components.

Publishers can be considered first-party potential attackers within the online advertising ecosystem. Attracting users to its web pages, a publisher receives direct requests from them. From such requests, some items of user information can be immediately inferred such as location and agent string. Depending on the type of publisher (news, shopping, social network, rating, etc.), certain information about the user such as gender, age, shopping habits or preference ratings may also be collected. The tracking mechanisms used by publishers are supported on their web log files and first-party cookies.

Advertisers become third-party adversaries since they receive information about users from subtle requests that derive from a user's page visits. Browsing history, location, gender, shopping habits, and other basic contextual data is typically leaked by the online advertising infrastructure so that advertisers can decide whether to bid or not for a given user impression. However, since the described interaction is currently subcontracted to aggregating entities like DSPs and ad networks, the ability of advertisers to directly access user information is significantly diminished.

The ability of **DSPs** to aggregate user information make these intermediaries very powerful potential adversaries to user privacy. Working for thousands of advertisers, a DSP is responsible for selecting the best impressions to bid on. This bidding process is carried out on the basis of both users' metadata and advertisers' specific campaign requirements. Users' contextual data are included in billions of bid requests sent by dozens of associated ad exchanges. Hence, it is difficult to imagine the amount of user information that DSPs are fed with, even without winning auctions. In fact, although ad exchanges recommend not to misuse the contextual information contained in such bid requests, a massive surveillance engine could be deployed through a group of colluding DSPs.

SSPs are the primary source of user information in the current automatic advertising architecture. Helping thousands of publishers interact with other intermediaries such as ad exchanges, SSPs make an offer of an ad space to at least one ad exchange when a user triggers an impression. To give context to such an offer, it is sent along with user data that SSPs gather from different sources. These data may include the visited website, cookies, and browsing information. Thus, SSPs consolidate huge amounts of user data, which raises serious privacy concerns, especially when much of this information comes directly from publishers. From a user's perspective, DSPs and SSPs are third-party adversaries, as they are fed with private, sensitive information that does not come directly from users.

Acting as gateways between buyers (DSPs) and sellers (SSPs), **ad exchanges** are one of the strongest third-party adversaries in our privacy attacker model. These higher-level entities consolidate ad spaces offered by multiple publishers (SSPs) and organize automatic auctions to sell such spaces to advertisers (DSPs). With that objective, ad exchanges concentrate most of the online advertising traffic and the user information used as input to effectively distribute ads. But not only that, ad exchanges also massively distribute such user data to multiple advertisers (mainly DSPs) so that the latter can make their bidding decisions. Given such capabilities of consolidating and indiscriminately distributing user information, ad exchanges are clearly the most powerful privacy attackers of the online advertising ecosystem.

Finally, although they are not strictly part of the online advertising architecture, **broadband providers** are unsurprisingly part of the attacker model we have described. Offering the transport channel that connects every user with the Web, these network-layer intermediaries have privileged access to user information, including that of ad related interactions. Table 3.1 consolidates the key details of these elements that make up the adversary model in the online advertising ecosystem.

3.2.2 Classification of Privacy Threats and User Role

Having specified the adversary model assumed for online advertising, which we described on the basis of the different intermediary entities involved in the ad-delivery

Component	Attacker's role	User collected data	Scope	Aggregation ability level	Privacy risk level
Publisher	First-party	clickstream, local browsing history, preferences, demographics, agent string, identification	Local	Low	Low
Advertiser	Third-party	restricted browsing history, preferences, demographics, identification	Local/Global	Low	Medium
SSP	Third-party	clickstream, restricted browsing history, preferences, demographics, agent string, identification	Global	Medium	High
DSP	Third-party	restricted browsing history, preferences, demographics, identification	Global	Medium	Medium
Ad exchange	Third-party	clickstream, detailed browsing history, preferences, demographics, agent string, identification	Global	High	High
Broadband provider	First-party	every single trace of user interactions with the Web	Global	High	High

Table 3.1: Components of our adversary model in the scenario of online advertising.

process, next we proceed to classify the corresponding privacy threats based on the capabilities of such entities, but also in terms of the limitations of users.

Platform Intrinsic Leaks

The main cause of privacy threats in online advertising is tightly coupled with the infrastructure and capabilities of ad platforms. To start, within this infrastructure, every tracking mechanism is enabled by default; there is not a built-in option for users to disable tracking or ad serving. Additionally, as depicted in Sec. 2.3, this infrastructure is significantly crowded with intermediate entities directly or indirectly fed with user data. Also, it is evident that the business model of online advertising, and so its infrastructure, builds on the collection of as much information about users as possible.

Regarding their capabilities, online advertising platforms carry out practices that support advanced levels of user targeting while neglecting privacy and even supporting the leak of personal data. In this subsection, we briefly examine such practices, which are mainly based on user tracking [57, 75]. Based on the interaction among users

Code	Privacy threat	Brief description
T1	First-party tracking	user information leaks out directly from the user side to the publisher
T2	Third-party tracking	user information leaks out from interactions between intermediate advertising entities and the user
T3	Cookie matching	user cookies are mapped and shared between ad exchanges and advertisers
T4	Fingerprinting	an identifying agent string is derived by first and third parties from certain specific characteristics of user applications and devices
T5	Flash cookies	intrusive and persistent cookie technology enabled by Flash-based websites
T6	Canvas fingerprinting	enables user tracking based on a fingerprint generated by the rendering of Canvas HTML5 elements
T7	HTML5 local storage	long persistent cookie-based tracking technology developed as part of the HTML5 language

Table 3.2: Summary of the privacy threats examined in our analysis.

and privacy attackers, tracking mechanisms can be classified into first and third-party mechanisms. As we see next, these mostly employ cookies to individuate users. Table 3.2 summarizes these threats.

T1. First-Party Tracking encompasses the activities performed by first-party adversaries, mainly publishers, to collect and analyze user information. Such activities include serving (first-party) cookies directly by the publisher to its users and mining the firsthand information, mainly location and agent string, provided by them in their web requests. Depending on the publisher's interaction level with its users, very valuable personal information could be directly gathered by publishers, e.g., gender, ratings, social interactions, preferences, shopping habits, health condition. Since the interactions leaking this information are explicitly triggered by the user, they are unlikely to be cataloged as malicious. Thus, detecting or blocking first-party tracking is just as complex, yet the scope of first-party tracking (and thus its privacy risks) is limited due to the size of the publisher's audience. Though, some publishers might collude with aggregating entities such as ad exchanges to provide them with aggregated user information [76].

T2. Third-Party Tracking builds on indirect, and likely non-consented, interactions between intermediate advertising entities (DSPs, SSPs, ad exchanges) and users. Such interactions are generated by content embedded in first-party sites from which user information is also leaked to third parties. The wider scope and higher

hierarchy of entities performing third-party tracking for digital advertising facilitate massive aggregation of personal information. However, third party tracking is not only deployed through cookies, but also by means of social plug-ins that may also disclose user browsing information to social networks [77]. Mechanisms aimed at protecting users from privacy risks of online advertising commonly block third-party connections after classifying them as undesired [78].

T3. Cookie Matching is a technology that supports the sharing of user data. Served both by first and third-party adversaries, cookies are the basic tracking technology used in online advertising. Within online advertising, cookies have given rise to concerns about the privacy of users for two main reasons. First, cookies are currently being used to store personal information such as e-mail addresses, not only identifiers to recognize a user in future visits [45]. Secondly, they enable massive sharing of such personal data through a more refined tracking technology, CM. CM enables an ad exchange to share users' cookie information with multiple potential advertisers so that they can infer contextual user data by mapping their own cookies (obtained from previous interactions with a user) with the ones obtained from the ad exchange [44].

Experiments done by Bashir et al. in [76] report about the ubiquity of CM on today's Web and on how shared information supports highly targeted advertising. It is worth noting that, although using cookies is an old practice originally built upon pretty small pieces of identifying information, they have significantly evolved to become large capacity structures, very popular tracking mechanisms, and increasingly more difficult to delete, as illustrated in Tables 3.3 and 3.4. Accordingly, a great deal of recent research has been done regarding online tracking [79–81], studied in desktop browsing contexts where the most evolved forms of cookies [82, 83] are subject to analysis.

T4. Fingerprinting, not built on cookies, is also available to support personalized online advertising. It consists in detecting the agent string of users' devices or applications. Thus, no matter if a user deletes her cookies, he can always be tracked online through such an agent string [46]. As a matter of fact, some variations of fingerprinting are commonly used to respawn cookies after a user deleted them. Mayer

	Max. storage size	Level of persistence	Storage location	Difficulty to delete	Usage level	Installation	Access level
HTTP cookies	4 KB	low	within the browser	low	remaining	native	one browser
Flash cookies	100 KB	medium	outside the browser	high	declining	through a plug-in	multiple browsers
HTML 5 cookies	5 MB	high	within the browser	high	increasing	native	one browser

Table 3.3: Comparison of the types of cookies that are typically used to track users.

	Effectiveness individuating users	Ad companies involved	Have led to lawsuits?	Easily erasable from browser?	Usage level	Are intrusive?
HTTP cookies	High	All [79]	No	Yes	Extended	No
Flash cookies	High	`hulu.com`, `about.com`, `aol.com`, Clearspring, Interclick, Quantcast [40] [79]	Yes	No	Extended	Yes
Canvas fingerprinting	Low	Addthis [79]	Yes	No	Limited	Yes
HTML5 local storage	High	Ringleader Digital, Bluecava [40] [83]	Yes	No	Growing	Yes

Table 3.4: Tracking mechanisms used in modern online advertising.

and Mitchel synthesize in [40] a list of non-cookie web tracking technologies used both from first and third-party entities.

T5. Flash Cookies [82] pose an alternative tracking technology for advertising entities trying to face the advent of mechanisms to block traditional tracking. Flash cookies are more effective in tracking users than common HTTP cookies. In fact, Flash cookies are considered prominently intrusive due to their persistence characteristics (more storage capacity, browser independent storage, and non-default expiration) [82–84]. After online advertisers were accused of misusing Flash cookies (by enabling restoring of deleted HTTP cookies), a study by McDonald and Cranor [85] found that the practice of respawning erased cookies had become significantly less aggressive.

T6. Canvas Fingerprinting is another persistent web tracking technology currently used by some online advertising agents, especially data aggregators [86]. Canvas fingerprinting facilitates tracking by generating a fingerprint of a user's browser from an HTML 5 Canvas element [79]. Such an element might be used by an (first or third-party) adversary to dynamically display, even invisible, text or images in the user's browser. Since the rendering of the Canvas element will slightly vary depending on the web browser's image processing resources, such particular displaying parameters could be used to get a fingerprint that might uniquely identify a user surfing a web page; to do it, certain browser properties are collected such as the list of installed plug-ins [46]. A few first and third-party providers of Canvas fingerprinting have been found from previous studies [79] and the tracking mechanism can be blocked if the provider's domain is known.

T7. HTML5 Local Storage is an even more persistent cookie-based tracking technology, developed as part of the HTML5 web language. Local storage enables more universal user tracking [87] that does not depend on the browser used, does not expire, and offers even more storage capacity, by default, than HTTP and Flash cookies (see Table 3.3). Such a feature might let some first or third parties store data (within the user's browser) that cannot be deleted when erasing browser's cookies. However, such intrusive tracking mechanisms might be aggressively tackled with lawsuits, especially when accomplished by advertisers, as Wired reported in 2010 [88]. Said misusing of cookies was reported by Hoofnagle et al. in 2012 [87] when they found that some companies had been using HTML5 and Flash cookies to respawn HTTP cookies that had been previously deleted by users. In Table 3.4 we summarize some of the characteristics of these tracking mechanisms including their effectiveness in individuating users, and whether the companies using them have faced lawsuits due to the intrusiveness of these mechanisms.

Other intrinsic properties of ad platforms make them pretty susceptible to privacy leaks. For example, the subtlety of their background processes isolates users in a separate dimension where they are unaware of the implicit risks. In addition, as recently reported in [45], relevant user information might be being conveyed in the

clear text during real-time auctions. In the same line, [14] and [76] reported cooperation between relevant entities such as ad exchanges and publishers, and quantified the derived leakage of users' browsing information. On a last note, chances are that the context information that feeds auctions will reach entities not really involved in bidding processes (or deliberately bidding to lose). Should ad platforms cannot detect such behavior, a cheap massive surveillance tool could be built on top of advertising infrastructures. This phenomenon is experimentally addressed later in Chapter 5.

User Role Limitations

User capabilities are, by default, pretty limited online. Although their interactions fuel ad delivery services, users are unaware of the transactions that are made in the background when they are served an ad, which also reduces their chances to protect themselves. This blindness and lack of control of users is the source of important privacy threats, especially in online advertising systems, where ad services are inherent to web browsing.

L1. Lack of awareness. Historically, online privacy has been a concern for users, as reflected in [89]. However, as explained by Ackerman et al., when faced with an abstract context where the leakage of personal information is not evident (as it might be within social networks), users' concerns get significantly lightened. This attitude of users towards privacy, particularly in advertising environments, is illustrated in [90], which reports that users are more concerned about being shown embarrassing ads than about being tracked.

In accordance with said lack of awareness, users hardly notice the relative value of their data within commercial contexts. Evidence on the dichotomy on how users and ad services value user data is offered in [91] and [14], respectively.

L2. Lack of control. In the opaque scenario of online advertising, users cannot protect their privacy adequately. Neither their interests nor concerns can be enforced because users are, by default, passive entities in the advertising ecosystem.

L3. Bounded technical knowledge. Users face an important cognitive barrier that seriously limits their capabilities to manage their protection against privacy threats in online advertising. Even being aware of the risks posed in this context,

and having the control to at least mitigate some of them, most users do not have the technical knowledge to understand the logic of protecting themselves within such a complex scenario.

Consequently, in online advertising contexts — unlike what happens in other online scenarios —, leaks of user data are not driven by user explicit flaws but arise from the complex structure and operation of the ad-serving process. Ironically, online advertising was said to offer users more control over advertising exposure than traditional advertising [92]. Table 3.5 summarizes the user role limitations that exacerbate user privacy risks in the context of online advertising.

3.2.3 Impact of Online Advertising Practices on Privacy

Since ad personalization (e.g., based on location, context and interests) increases conversion rates, users' browsing data have inevitably become an asset that nowadays is exchanged throughout the entire online advertising infrastructure [45]. The need to further scrutinize this information to profile and segment users raises serious privacy concerns with respect to social sorting and discrimination, particularly as potentially sensitive information can be inferred from the profile of a reidentified user, such as income level, health issues or political preferences.

Modern auction-based ad delivery requires that processes be executed in real-time, which implies that vast amounts of user information be mined at very high rates. This urgent need might naturally discourage the online actors from protecting user information against privacy attacks. Besides the urgency in which data must be handled, the need to offer tailored ads compels the advertising ecosystem to collect a wide range of metadata. For this reason, practices such as cooperation (collusion) among advertising entities and aggregation are enabled to facilitate massive and often uncontrolled sharing of said information [44]. Since the shared data (sometimes including even the prices paid by advertisers) are not always encrypted, other adversaries, such as Internet providers, come into the picture.

As described in previous sections, online advertising builds on non-transparent interactions among a myriad of intermediary ad companies, which have the ability to

Code	User role limitations	Brief description
L1	Lack of awareness	the leakage of personal information is not evident for users in online advertising
L2	Lack of control	user preferences and concerns are not technically enforced by default in online advertising
L3	Bounded technical knowledge	users barely have the technical knowledge to understand and effectively use protection tools

Table 3.5: Summary of the user role limitations examined in our analysis.

profile web users. As a result, not even publishers are aware of which information is collected and how it is used. In fact, publishers are unaware of what ads are shown to their visiting users. The ad-delivery process involves so many intermediary companies that it is impossible for an ad exchange to control the use of user data by such companies. In fact, cases are known where attackers took advantage of advertising channels to distribute malicious code to millions of users [6]. This lack of transparency obviously prevents users from actively getting involved in the protection of their privacy. Though there are informed users who use transparency and protection tools while browsing, advanced mechanisms are currently implemented by the online advertising ecosystem to counteract cookie removal or ad blocking.

Finally, due to the auction-based policies of the advertising ecosystem, certain users invariably become more economically valuable than others. For example, Olejnik et al. found in [14] that, in terms of prices paid during online auctions, visitors of websites belonging to particular categories are much more relevant than visitors of websites of other categories. Yet, other criteria such as the user location and time of visit might also be used to determine the relevance of the corresponding profiles. Such more relevant users stand out from the rest and gradually their profiles become more identifiable and, as a result, less private. Unfortunately, evidence has been found suggesting that negative discrimination (such as racism) might be performed in online ad delivery [93].

3.3 Privacy Protection Approaches

In this section we perform a deep analysis of the privacy protection approaches available for users in the context of online advertising. For the sake of comparability, we

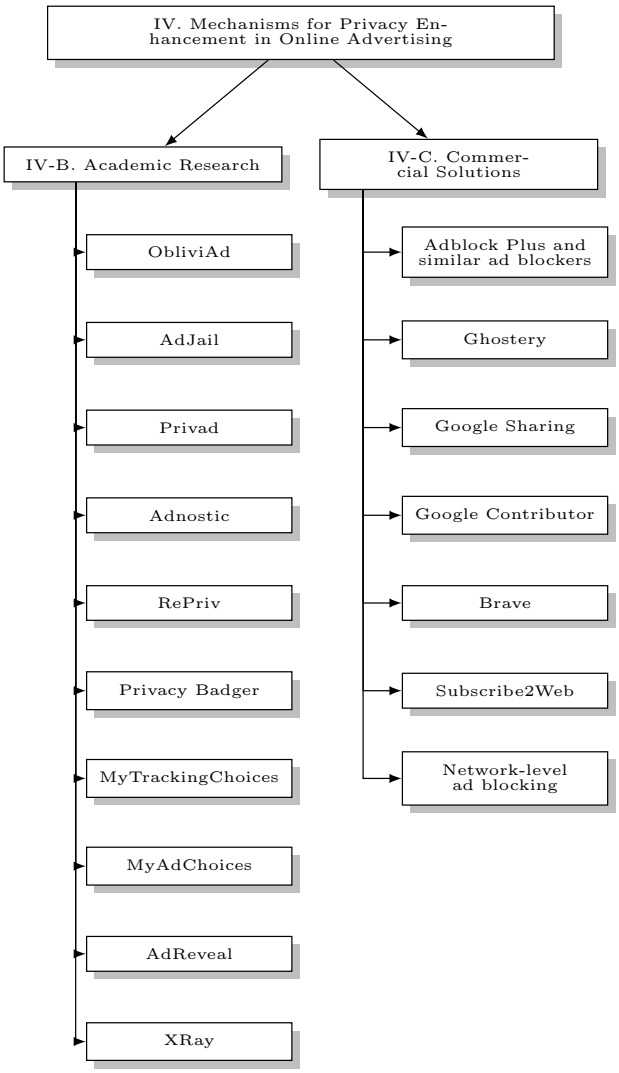


Figure 3.1: List of privacy mechanisms, specifically intended for online advertising, that we examine in Sec. 3.3.

first define the protection parameters based on which each mechanism, tool, or technology is described. Moreover, we classify them as academic or commercial initiatives since many of them have been implemented and are very popular among users. For illustration, the full list is of these approaches is presented in Fig. 3.1.

3.3.1 Protection Parameters

Our analysis of privacy protection mechanisms examines three main characteristics of their operation: location, scope, and strategy. Below we proceed to describe them.

Location

According to the location where the protection mechanism takes place, the current research proposals and commercial solutions can be classified roughly into local and third-party. On the one hand, local mechanisms commonly lie on the user side, for example, in the form of an application running on the user's browser, or as a local service operating in the user's network [53]. Some academic approaches propose migrating the profiling processes required for ad targeting to the user side [57]. On the other hand, third-party mechanisms are implemented with the help of a broker entity, whose location is remote from the user side, and whose aim is commonly to provide security services such as secure storage of data, anonymization and even user profiling [51]. We would like to stress that, even in the case of broker-based mechanisms, a local application on the user side is frequently required to engage users with said broker.

Scope

Depending on the scope of application of the mechanism in question, we may characterize it as local or global. Protection approaches whose scope is local usually aim at adapting a protection mechanism to the structure of the current advertising ecosystem. Hence, the scope of protection offered is also limited to the information and interactions available to the user. On the other hand, those protection approaches with a global scope come in hand with new ad delivery models, pretending to radically change the manner in which ad serving processes currently function, especially with regard to their relationship with users. The majority of these approaches has been envisioned as privacy-by-design models of advertising which would provide users with significant control over their interactions with ad platforms.

Strategy

In our classification of privacy technologies, we also consider the principle or *strategy* that rely on. We contemplate five strategies which range from user lack of awareness through *transparency*, to undesired interactions with third-parties by means of *blocking*, *obfuscation*, and *sandboxing*, and to a by-default exclusion of users from the advertising logic through more *inclusive* techniques. Next, we describe these strategies.

Transparency: Undoubtedly, a first step towards privacy protection may be *transparency*. Transparency in this context means allowing users to learn what is going on with regard to their activity and data in online advertising systems. Some of the approaches examined in the coming subsections provide transparency usually by making users aware of the tracking activities behind the scenes, and by allowing them to know how their browsing traces might have been exploited to deliver targeted ads.

Blocking: Blocking is also a very common, although usually radical, strategy of privacy protection in online advertising [78]. Typically, blocking tools inhibit most of the known tracking mechanisms (and thus of advertising) from the user side, or a third-party located on their network. Because the vast majority of ads are delivered nowadays through third-party trackers, cutting of third-party tracking implies eliminating nearly all ads. Originally, blocking mechanisms had been designed as a binary choice, namely, either blocking or allowing all tracking and hence advertising. Nevertheless, recent academic proposals tend to lighten this radical strategy by providing fine-grained control over tracking, by enforcing users' preferences and by using smart and dynamic learning mechanisms [94] [95].

Obfuscation: It consists in perturbing sensitive data in order to preclude an adversary from discovering the identity of its owner and/or deriving private information about them [96]. In the context of online advertising, some privacy protection approaches implement obfuscation by mixing data and metadata of a group of user profiles so that the intrinsic features of individual profiles cannot be recognized. Other approaches build on external brokers to anonymize user data by randomly masking potentially identifying attributes such as IP addresses and cookies.

Protection mech.	Location	Scope	Protection strategy				
			Transparency	Blocking	Obfuscation	Sandboxing	User inclusion
ObliviAd	local, third-party	local			✓	✓	
AdJail	local, third-party	–		✓		✓	
Privad	local, third-party	local			✓	✓	
Adnostic	local	local, global			✓	✓	✓
RePriv	local	local		✓		✓	✓
Privacy Badger	local	local	✓	✓			✓
MyTrackingChoices	local	local	✓	✓			✓
MyAdChoices	local	local	✓	✓			✓
AdReveal	local	local	✓				
XRay	local	local	✓				

Table 3.6: Academic proposals for privacy enhancement in online advertising classified on the basis of the protection parameters described in Sec. 3.3.2.

Sandboxing: Sandboxing addresses security threats by isolating suspicious applications from the resources they rely on. Within online advertising, sandboxing is applied by keeping apart certain critical processes which may give advertising brokers access to sensitive user data. A typical sandboxing example leverages on the execution user profiling on the premises of the user, rather than on the ad-platform side [52, 57, 58].

User Inclusion: With the aim of balancing the Internet's dominant business model and user privacy, some proposals envision a more user-driven ecosystem. In general, giving users more control over their interactions with ad platforms might help achieve said balance. A practical step towards this consists in adapting the protection mechanisms to the needs of users. In this line, most ad blocking solutions have recently started to offer users some personalization features such as blocking per domain and per tracker. Other strategies include the enforcement of user choices over third-party, cookie-based tracking, and the provision of direct interfaces with the advertising ecosystem [58, 97].

3.3.2 Academic Initiatives

This section examines in depth the most relevant approaches in the academic literature of privacy-enhancing technologies for online advertising. Table 3.6 provides an overview of the technologies explored in this section.

ObliviAd

Proposed by Backes et al. [51], ObliviAd relies on adapting secure-coprocessor-based brokers to the current online advertising ecosystem. The aim of such coprocessors is providing private information retrieval of user data during the delivery of ads to users and the billing to advertisers. To do so, this approach provides different services such as the secure storage of sensitive user data, the encryption of profile information when it is conveyed to the broker side, the encryption of ad information to be displayed on the user side, and finally the obfuscation of billing data to charge advertisers.

While all these services that integrate ObliviAd may offer strong security guarantees through hardware and heavy cryptographic techniques, this is undoubtedly at the cost of a significant increase in complexity and deployment. It is worth stressing as well that network and browser identifiers such as the user's IP address and user agents might still leak, which means that this approach might not be useful against the fingerprinting techniques described in Sec. 3.2.2.

AdJail

Ter Louw et al. [98] proposes a tool that aims at empowering publishers to isolate the content elements to which ads will have access to. Specifically, this approach allows safeguarding a user's scope and that of the web application by creating a sandbox where ads are executed. From this sandbox, ads may have access to user or publisher content through a configurable set of enforcing policies. Although the aim of AdJail is to protect the confidentiality and integrity of user and publisher data, user privacy can also be provided by applying those policies based on the privacy agreement negotiated between publishers and their users.

The problem of AdJail, however, is that its scope is limited to the publisher's domain. In other words, users can utilize this sandboxing approach only if this mechanism is deployed in the website. In addition, AdJail focuses more on other security services such as integrity and confidentiality, and does not tackle the privacy threats identified in Sec. 3.2. Since AdJail's scope focuses on the publisher, we categorize this proposal neither as local nor as global.

Privad

S. Guha et al. [52] seeks a more private online advertising system and offers to this end an alternative private solution that may adapt seamlessly to the current advertising business model. The authors argue that Privad, their solution, would preserve privacy by keeping a user's browsing profile within a local user application. Nonetheless, they also claim that some information (related to the user's interests and to the ads they have viewed or clicked) "necessarily" would have to leave the user's domain.

Privad also incorporates a third-party anonymizing proxy. This proxy would receive the released (and ciphered) user information and, after hiding the user network address, it would deliver this information to an ad platform. Advertisers aiming at delivering ads feed the ad platform with their ads, including information of the profile to which each ad is targeted; and then this information is employed by the ad platform to tailor ads to those profiles. Consequently, this approach uses the proxy to anonymize user information so that the ad platform in question is not able to individuate a user from the preferences reports generated by his browsing activity. Unfortunately, anonymizing strategies like this have proved to be weak [74], especially when demographic information about users is still available for a potential attacker, and when such information is managed by a third-party entity over which a user might not have any control (such as an Internet service provider).

Adnostic

It is an academic proposal by Toubiana et al. [99] that implements a more friendly architecture to display personalized advertising without compromising user privacy. Such architecture does not rely on blocking ads but on performing the whole user profiling process within the user domain, so that no personal information is leaked out to third parties.

The ads to be shown to a user are chosen on their side, according to a locally estimated browsing profile. This profile is constructed by processing the user's queries and the content of visited pages. Then, this information is classified within the

browser by means of natural language processing techniques. The ads, which are part of a previously downloaded set, are displayed according to the user's interests.

Because personalization is not directly controlled by ad platforms, there are less incentives for advertisers to bid more money to place ads. However, we may expect worse personalization performance since this process takes place on the user side, based only on their browsing data. This is in contrast to the current ad-targeting algorithms implemented by ad platforms which rely on massive amounts of aggregated user data.

In terms of impact on the current infrastructure, on the other hand, Adnostic would eliminate the requirement of intermediary ad platforms, but unfortunately at the expense of less effective ad-targeting. As a matter of fact, the more components of the online advertising architecture are embedded on the user side, the more control the user may have over advertising. Obviously, this would mitigate many of the privacy risks analyzed in Sec. 3.2.

RePriv

It is a proposal by Fredrikson and Livshits [58] that aims at carrying out a selective disclosure of user data through a browser-based tool. First, as with the extensions described above, the proposed system would rely on the ability of the browser to capture all the information spawned while browsing the Web; this is the basis for local user profiling. Next, the system contemplates that the interests derived from such user profile are released to third-parties *only if the user gives permission*. Detailed information about their browsing habits, though, would not be released by default. Finally, the proposed system considers additional modules that would interface with third-party applications interested in having access to user data.

The privacy-preserving strategy of RePriv consists in profiling users locally, so that they have control over the information that is disclosed to ad companies. However, although users are in control of said disclosure, external entities might be collecting such data anyway. Even though at first RePriv might seem an interesting approach, its success in protecting user privacy certainly depends on the disclosure control given

to users. Again, such a control may tend to be absolute (as in ad blockers) or could be softened to balance the trade-off between user privacy and the web business model.

Privacy Badger

Much of its functionality was incorporated from an older project called ShareMeNot which was originally presented in [100] by Roesner et al. Currently supported by the Electronic Frontier Foundation (EFF), Privacy Badger is an open-source browser extension developed for Chrome and Firefox [94]. The extension was not conceived as an ad blocker but as a privacy tool that may prevent non-consented tracking.

The operation of this browser plug-in does not rely on blocking all tracking by default and on static filtering lists. Instead, it capitalizes on an algorithm to detect and then prevent non-consensual tracking activities. Since the blocking mechanism is not based on the subscription to a deliberate filtering list but on rigorous algorithmic methods and policies, engagements with advertising companies to include blocking exceptions are in principle less likely to occur.

With regard to its graphical interface, this extension looks very similar to Disconnect and Ghostery. The user is shown the tracking companies following their visit to a page. As mentioned above, this tool does not block a tracker unless its algorithm checks it is following the user without their consent. Nevertheless, conducting this checking may take some time. On the other hand, as with most ad blockers, users may individually block or allow the detected trackers, or block only the corresponding tracking cookies. Additional options include disabling the extension on a per-site basis and manually adding a whitelisted trackers domain.

Privacy Badger represents a promising approach to balance the trade-off inherent in online advertising between user privacy and the Web economic model. In fact, besides blocking non-consensual tracking, its developers offer ad companies the opportunity to be whitelisted if they formally promise to respect opt-out mechanisms (e.g., Do Not Track headers), conforming with users' privacy policies [94].

MyTrackingChoices

Achara et al. [97] propose a browser extension available for Google Chrome and Mozilla Firefox. The plug-in targets users who are not in general against advertising and accept the trade-off that comes with the "free" content. However, for privacy concerns, they wish to exert fine-grained control over tracking.

This academic proposal relies on the assumption that some categories of web pages (e.g., related to health or religion) are more privacy-sensitive to users than others (e.g., about education or science). Based on this idea, the plug-in allows users to specify the categories of web pages that are privacy-sensitive to them and block the trackers present on such web pages only. As tracking is prevented by blocking network connections of third-party domains, MyTrackingChoices avoids not only tracking but also third-party ads.

The detection of the tracking companies does not rely on existing blacklists, unlike most ad blockers and anti-trackers. Rather, MyTrackingChoices keeps a local list that is built from the pages browsed by the user. This list is smaller and easier to maintain than the list of tracking and advertising domains currently used by Adblock Plus. To decide if a third-party domain is a tracker or not, the tool checks it is present on three or more different domains that a user visited in the past. Since users continue receiving ads on those web pages which belong to non-sensitive categories, this approach may provide a better trade-off between user privacy and the Web economy. However, this approach only provides privacy protection against previously defined sensitive content, when tracked through HTML cookies, and thus does not preclude more sophisticated tracking technologies (such as canvas fingerprinting) and less simple tracking methods based on IP address.

MyAdChoices

Parra-Arnau et al. [95] propose a web-browser plug-in aimed at bringing transparency over tracking and advertising, and providing a certain level of granularity with regard to blocking ads. As for transparency, the plug-in estimates if the ads delivered to a user may have been generated from their previously visited pages. It also permits

users to know if the browsing profiles available to trackers and ad companies may show common or unique interests.

In terms of blocking functionalities, the tool enables users to hide ads by topic category and depending on whether they have been displayed based on users' browsing interests or not. Although the tool provides fine-grained control over ads, it does not prevent any form of tracking; ads are basically hidden to users by applying a black mask on top of ad images. Another limitation of this approach is that the transparency functionalities come at the cost of additional traffic. The reason is due to the fact that, to decide if an ad is profile based, it must revisit the pages browsed by the user in incognito mode.

AdReveal

Liu et al. [101] propose an advertising-transparency platform aimed at studying the ads delivered to some artificial profiles, built from the AOL search query data set [102]. The tool is not intended for end-users, unlike MyAdChoices, and provides a framework that aims to study interest-based and contextual advertising at large scale. The platform, which operates offline and is restricted to DoubleClick ads, analyzes two data sets to this end: the interest categories of *all* ads received both in a tracked session and in an incognito-browsing mode. The authors then use a binary classifier to decide if an ad belonging to a certain category is interest-based *or* contextual.

XRay

Similarly to AdReveal, XRay [103] propose a transparency platform which tracks the personal data collected by several Web services, and tries to correlate data inputs (e.g., e-mails and search queries) with data outputs (e.g., ads and recommended links). The proposed platform has been tested for the ads displayed on Gmail and relies on the maintenance of a number of *shadow accounts*, that is, replicates of the original account (e.g., an e-mail account), but which differ in a subset of inputs. All these account instances are operated in parallel by the system and are used to compare the outputs received. Intuitively, if an ad is displayed more frequently on those accounts

sharing a certain input (e.g., an e-mail), and this ad never shows up in the rest of shadow instances, then this input is likely to be the cause of said ad.

The major limitations of transparency tools such as MyAdChoices, AdReveal and XRay, come from the necessarily simplified model assumed for the ad-delivery process. Evaluating an ad-transparency tool is, besides, extremely challenging since the ground truth of targeting decisions is unknown. XRay, in addition, provide a solution which is not intended for end-users, i.e., it is not designed to be used by a single user who wishes to find out what particular ads are targeted to them.

3.3.3 Industrial and Commercial Solutions

Commercial solutions mostly take the form of web-browser extensions. Since all user interactions with the Web are handled through the browser, taking advantage of such an interface to filter or block third-party tracking seems a reasonable approach. These browser extensions endeavor to protect user privacy by blocking third-party interactions. This strategy is usually implemented both statically, based on lists of banned trackers, or dynamically, based on heuristics and automatic learning. The specific implemented approach has an immediate, evident effect on the trust level over the tool and even the performance of the browser. For instance, those tools based on large blocking lists (such as the ones available for the most popular browser extensions) may perform worse due to the need to check these lists every time a page is visited. In this regard, we hasten to stress that the criteria employed to manage such lists is not clear at all. This obviously may arise suspicion and reduce the level of trust in these solutions.

A rich variety of browser-based solutions are currently available as commercial products, some of them providing users with control over online advertising. The controversy stirred by the use of the blocking lists they rely on [104] [105], however, has motivated the rise of *open-source, transparency* technologies that may prevent ad companies from interfering. Next, we examine a particular class of solutions called ad blockers. Although there exist numerous tools of this kind, our analysis will focus

Protection mech.	Location	Scope	Protection strategy				
			Transparency	Blocking	Obfuscation	Sandboxing	User inclusion
Adblock Plus and similar	local	local	✓	✓			
Ghostery	local	local	✓	✓			
Google Sharing	third-party	local		✓	✓		
Brave	N/A	global	✓	✓			✓
Subscribe2Web (2)	N/A	global		✓			✓
Google Contributor (2)	N/A	global		✓			✓
Network-level ad blocking	local, third-party	local	✓	✓			✓

Table 3.7: Summary of the commercial solutions for privacy protection in online advertising, classified according to the parameters examined in Sec. 3.3.1.

Extension	Blocking strategy	Trust level	Expected performance
Adblock Plus	List-based	Medium	Low
Ghostery	List-based	Low	Medium
AdBlock	List-based	Medium	Low
Disconnect	List-based	High	High
Lightbeam	List-based (items added manually)	High	Medium
Privacy Badger	Heuristic-based/dynamic	High	Medium
DoNotTrackMe/Blur	List-based	Medium	Medium
MyTrackingChoices	Dynamic	High	High
MyAdChoices	Dynamic	High	High
Brave	List-based	High	High

Table 3.8: Browser-based approaches described in terms of their blocking strategies, trust level and performance.

only on the most popular ones, namely Adblock Plus and Ghostery. Other ad blockers such as AbBlock [106], Lightbeam [107], Disconnect [1], Blur [108], SuperBlock Adblocker [109], AdRemover [110], AdBlock Pro [111] and uBlock [112] operate similarly.

The last group of (four) initiatives explored in this section aim at radically changing the paradigm of the online ad delivery. Sponsored by relevant institutions such as Google, Yahoo and Internet providers, these initiatives propose a user-driven architecture whose main aim is to strike a better trade-off between user privacy and the Web economic model. Table 3.7 shows a classification of the commercial solutions analyzed in the coming subsections on the basis of the protection parameters described in Sec. 3.3.1. Table 3.8 shows different aspects of the browser-based proposals (both academic and commercial) such as their strategy to prevent tracking, and the corresponding trust level and performance.

Adblock Plus

Available for all major browsers, Adblock Plus is an extension that blocks tracking and ad serving [47] based on filtering lists which specify the elements of a website that may be blocked. These elements include malware domains, banners, pop-up windows, and video ads on Facebook and YouTube. Users enable blocking by adding the filtering lists of their preference, managed in [54]. Adblock Plus is the world's most downloaded ad blocker and therefore the tool that is currently threatening the Internet business model [113].

This ad blocker has recently incorporated a whitelisting mechanism —enabled by default— for nonintrusive ads that meet certain criteria. These criteria are defined in the *acceptable ads* initiative [114], and although the adherence to this initiative is optional for advertisers, much criticism has arisen especially after the revelation that Adblock Plus was getting money from ad companies to whitelist them [115, 116].

Ghostery

Developed by Evidon, Ghostery [117] is a proprietary browser add-on capable of detecting third-party trackers. By default, this tool blocks the execution of the tracking cookies as well as the scripts belonging to the tracking companies that are blacklisted. The list in question is elaborated by the company itself. Even though the tracking companies in this list are classified into five categories, according to their different purposes (analytics, web bugs, privacy, advertising, and widgets), it is highly unlikely that users recognize such categories or entities to make a conscious configuration of the tool [118]. However, using such lists may simplify the configuration of the add-on.

When a user browses the Web, Ghostery shows the trackers that are blocked on each page (through a non-intuitive or usable categorization), and offers the possibility of adding any such trackers to a whitelist. Ghostery protects users' privacy from advertisers by blocking scripts, images, objects, and documents embedded by companies the user might not trust. Other tracking mechanisms such as web or canvas fingerprinting are not addressed by Ghostery. Finally, the tool has been criticized

for its default behavior [118] which allows Ghostery to collect information about the blocked ads, and afterwards sell it to ad companies [119].

Google Sharing

Google Sharing [120] is a system that provides privacy protection by avoiding the tracking conducted by Google. It consists in a Firefox extension that redirects user's requests to an external proxy, where a group of identities associated with cookies are managed. These cookies replace the ones included in original requests, masking a user's identity, and are then forwarded to Google along with the original request. Even when they allow users to send encrypted requests, user privacy can still be compromised if collusion exists between the proxy server and Google servers.

Brave

It is a web browser —and not a plug-in— that natively embeds functions to block intrusive ads and third-party tracking by default [62]. This proposal allows replacing the ads available on the visited pages with others from Brave's own advertising network, claimed to be less intrusive and more privacy-friendly.

The proposed browser contemplates integrating users into the online advertising business by paying them 15% of the gross ad revenue. In this regard, users are given the option to donate such money to publishers, in exchange for an ad-free browsing experience. Among other transparency functionalities, users may learn the number and type of blocked ads, the trackers present on the visited pages and HTTP redirections.

The upshot is that Brave operates similarly to an entire ad platform, but managed by a single company. The solution completely dispenses with the present advertising infrastructure and aims at building a new one, apparently fairer and more private. However, this approach has sparked much criticism [121] since users' browsing data are collected and processed by a *single* company, which merely shift users' trust from the current multi-system advertising model to this new single entity.

Subscribe2Web

Developed by Mozilla, Subscribe2Web [63] endeavored to address some of the privacy risks examined in Sec. 3.2. Based on the idea that online advertising is crucial for the present Web content model, Subscribe2Web looks for a way whereby the main actors (in particular, content creators and users) can meet and have a natural exchange of value. Mozilla's proposal is to eliminate the current Web dependency on ads, in order to fund the content creation by directly compensating content and service providers. The aim is to provide the Web with an API accessible from any browser through which users would pay a monthly subscription in exchange for accessing ad-free content.

Google Contributor

Contributor [122] was an initiative supported by Google to reduce the amount of ads delivered by its advertising services. Its main aim is not directly related to protect user privacy but to give users the possibility to eliminate ads from their favorite sites. Because advertisers would be partially excluded by this approach, users registered with this service would have to somewhat support the free ad sites by paying a monthly fee. Thus, Contributor relies on a novel idea where users are considered as active agents in the Web economic model.

Network-Level Ad Blocking

Recently, some Internet service providers have started to cooperate with ad companies to implement ad blocking technologies [123] [124]. This is the case of Three, an operator in the UK and Italy, which is working with Shine Technologies to deploy network-based ad blocking.

With these network-level ad blocking practices, a new powerful agent breaks into the online advertising ecosystem, stating that customers should have more control over the content displayed on their browsers, especially when they would be paying for every downloaded byte. Even though not much information is available about the blocking mechanisms to be used, the goal would not be to eliminate advertising but to give users more information (transparency) and the option to decide what to

block (control). In the long term, this approach may help protect user privacy, offer relevant and non-intrusive ads, and allow advertisers to take upon the data charges for downloaded ads.

3.4 Discussion

In Secs. 3.1 and 3.2 we made it clear that online advertising is a market where the exchanged goods are the users' data. Therefore, the multiple interactions among the entities of such a market might entail privacy risks for its users. Third-party entities from online ad platforms, such as DSPs, SSPs and ad exchanges, and many others offering a transport channel are especially responsible for the collection and aggregation of most of the user information employed as the raw material for their targeted ad delivery strategies.

The main concern of privacy advocates about online advertising is that the user information collected by intermediate entities might be employed to uniquely identify users or classify them in order to, for instance, discriminate their patterns of behavior. This risk is significantly worse due to the following factors specific to the online advertising ecosystem:

- most processes are performed in the background so the infrastructure is not transparent by default for users;
- user data is massively collected by several intermediary entities;
- the user data are necessarily distributed and processed at very high speeds due to the real-time requirements of advertising, which makes it difficult for ad companies to anonymize and protect such data right after their collection;
- cooperation is encouraged between intermediate entities in terms of data sharing;
- multiple items of information can be collected about users through non-consented interactions that are indirectly triggered from users;

- information about users along with processed metadata are commonly exchanged in an unencrypted form between ad serving entities; and

- advanced, resistant and intrusive tracking mechanisms are used to identify users online.

The inevitable consequence of the aforementioned procedures, supporting, in practice, the massive trade of user profiles, is the abusive and nonconsensual identification and classification of users [71] which in extreme cases might entail, for instance, discriminative treatment [93] when they receive online services. These factors of the online advertising ecosystem promote the development of advanced mechanisms to track users through the Web. Practices such as CM, flash cookie setting, canvas fingerprinting, and device fingerprinting in general are massively implemented [44,79,82,83] and sometimes become so intrusive that users are tracked even when some of such fingerprints have been deleted. Most of these practices build on cookies as a mechanism to identify users and to even store information about them. Cookies are commonly combined with other technologies such as canvas and device fingerprinting to obtain a less ephemeral trace of users. Meanwhile, CM exploits the identifying strings retrieved by using cookies to promote massive cooperation among online advertising entities.

The proved complexity of the online advertising ecosystem and the generalized control that huge companies have acquired over ad distribution infrastructures [11,25] significantly limits the scope of the proposed privacy protection policies. As a consequence, most of the privacy-protecting approaches build on local mechanisms which aim at disabling third-party interactions triggered from the user side to online advertising infrastructures (mainly between users and SSPs), directly blocking user information leakage. Such local approaches are commonly implemented as web browser extensions that provide users with transparency and ad control functionalities [47,95]. Still located between users and SSPs, other proposals suggest filtering strategies carried out by third-party entities (so-called brokers) [51–53] which may have access to the interactions directly performed between a group of users and the advertising

entities. Given the evident limitations of local approaches, some initiatives have envisioned privacy-by-design advertising platforms where privacy guarantees are provided with a global scope [62, 63, 122]. Interestingly, such initiatives agree on integrating users into the advertising ecosystem.

Our analysis has examined privacy mechanisms with various levels of impact on the Web. To start, offering transparency to users is probably the most appreciated feature of ad blockers (and research platforms such as AdReveal), which is complemented with tracking blocking capabilities to give users a significant level of control. Notwithstanding, the usability of ad blockers for nontechnical users is questionable [118] and these approaches dismiss much, if not all, the current online advertising ecosystem, thus hindering the current economic model of the Web supported by ads.

Even though some of these blocking-based solutions have become pretty popular (e.g., Adblock Plus), the changing business models and default (whitelisting) behaviors of some of these commercial solutions have stirred great controversy. Fortunately, other approaches supported by privacy activists, academics and foundations (such as EFF) are proposing more adequate and usable technologies (e.g., Privacy Badger, MyTrackingChoices) that may block tracking according to users' preferences [94, 95]. Other more refined variants of this blocking strategy are obfuscation and sandboxing [52, 57, 58, 98] (proposed by Obliviad, AdJail, Privad, Adnostic, RePriv and Ghostery). The ultimate aim of these mechanisms is also bounding the amount of user information learned by ad platforms, while striving to adapt to the current advertising business paradigm. As for the privacy threats posed by the structure and capabilities of online advertising, by blocking third-party tracking most commercial solutions claim to hamper cookie setting and thus CM. Canvas fingerprinting can be blocked by most local solutions, yet only on a per-domain basis, the same way as flash cookies. Remarkably, combining at least two ad blockers should offer enough protection against most of the threats described in Sec. 3.2.

Finally, given the dynamic nature of user and ad platform economic incentives [125, 126] with respect to privacy, it seems reasonable to propose new and more private ad distribution (and economic) models. Undoubtedly, this should be with the help of mechanisms that allows users to play a more active role on deciding whether to

be tracked or not [62, 63]. Inevitably, this level of control would imply an important reduction in revenue for publishers, and thus require users to directly pay content creators.

Since online privacy may be measured with respect to the interest of Web users to protect their browsing data and that of adversaries to exploit such information, analyzing the respective motivations of the different actors is also of great interest.

Without a doubt, economic incentives have encouraged intermediary entities, advertisers, publishers, and users to participate (consciously or not) in online advertising. Users' unconscious motivation to get involved in online advertising, playing the role of the product, is linked to their need to access free content and services on the Internet. Since the vast majority of Web content and services is paid from advertisement revenue, users have few options to opt-out.

On the other hand, publishers need to help advertisers and ad platforms in their bid to maximize their revenue. For this purpose, website owners are disposed to cede valuable space in their sites and information about their users to such intermediary parties, which thereafter will be responsible for deploying ad-delivery mechanisms. Thus, in exchange for money, publishers surrender some control of its interaction with users and indirectly participate in the disclosure of private contextual information to ad platforms.

On the other side, the interest of advertisers in actively leaking user information is rather reduced, unless several of them collude to share. However, advertisers typically engage ad exchanges and DSPs' services to receive contextual information, which may be useful to deliver targeted ads. Therefore, advertisers' incentives to collect user information are high as well.

The commercial nature of online advertising has spurred a debate about the motivation of the involved entities to protect privacy and to profit from user data. Although apparently opposed, the motivations of users and advertising intermediaries for privacy might vary according to factors that are not commonly considered. Research on the economic behavior of data holders in the market of online advertising [127] has shown that an increased level of user-targeting can reduce their profit due to an exacerbated transfer of value to advertisers. Specifically, advertisers would

be gradually less interested in bidding for user impressions as more detailed information is given to them. That way, according to Bergemann and Bonatti [126], an unexpected incentive may appear for data holders to provide reduced accuracy in the exchanged user data, with the aim of generating greater demand from advertisers and thus greater profit for data holders. Interestingly, such increase in profit may lead to more privacy for users (given by the reduced precision of user data leaked to advertisers). Nonetheless, a recent study by Taylor and Wagman [128] poses that the effects of targeting capabilities on profits depend on market and is, consequently, given by context.

Users seem to face a similar contextual dichotomy even though the concern about privacy is generalized [129, 130]. The fact is that the creation of a marketplace in personal data may shift the balance of power between individuals and companies that gather data. According to some recent studies, this is a shift people would be willing to embrace. Just over half of the 9 000 people surveyed worldwide said they would share data about themselves with companies in exchange for cash [131]. A separate survey has found that 42 percent of more than a thousand 13-17-year-olds in the U.K. would rather accept cash for their personal data than earn money from a job [132]. Lastly, it was reported in [133] that 56 percent of the consumers surveyed would be willing to give up personal data provided that they received some kind of economic compensation. This dichotomy between users' concerns and intentions regarding privacy might obey, according to Acquisiti et al. [125], to multidimensional factors relative to the context where the user operates, such as their lack of awareness about privacy risks, and cognitive and behavioral biases. The upshot is that users' assessment of their own privacy will strictly shape the impact of external threats. Tables 3.9 and 3.10 summarize our discussion of privacy technologies and how they may address the threats identified in Sec. 3.2.

3.5 Conclusions

Online advertising has become ubiquitous on the Internet and the revenues ad serving generates for publishers are supporting the existing free Internet access model. As

	Threats							User role			
	T1. First-party tracking (1)	**T2**. Third-party tracking	**T3**. Cookie matching	**T4**. Fingerprinting	**T5**. Flash cookies	**T6**. Canvas fingerprinting	**T7**. HTML5 local storage	**L1**. Lack of awareness	**L2**. Lack of control	**L3**. Bounded knowledge	**Observations**
ObliviAd		✓									Obfuscating user preferences may prevent third-party tracking. But IP address, user agents and other content embedded in websites might still be used as sources of fingerprinting
AdJail		N/A	N/A	N/A	N/A	N/A	N/A				Provides security services such as integrity and confidentiality in the publisher side
Privad		✓									May avoid third-party tracking, but other user data such IP address and certain user agent may be used as sources of fingerprinting
Adnostic		✓							✓	✓	If enforced by ad platforms, it would discourage third-party tracking. Protection against other threats is not considered
RePriv		N/A	N/A	N/A	N/A	N/A	N/A		✓	✓	Users may control their browsing data on their side, but nothing may prevent external tracking
Privacy Badger		✓	✓		✓	✓	✓	✓		✓	Blocks most tracking mechanisms, but little control is given to users
MyTrackingChoices		✓			✓			✓	✓	✓	Users may block third-party tracking on a more granular level, but protection is against previously defined sensitive content
MyAdChoices								✓	✓	✓	Does not prevent any form of tracking, and ads are hidden from the user, not blocked
AdReveal								✓			Framework aimed at studying interest-based and contextual advertising at large scale
XRay								✓			Platform, not intended for end users

Table 3.9: Online advertising privacy threats and academic proposals addressing them. (1) Since this threat derives directly from interactions explicitly triggered by users, protecting from it is a challenging task.

	Threats							User role			
	T1. First-party tracking (1)	**T2**. Third-party tracking	**T3**. Cookie matching	**T4**. Fingerprinting	**T5**. Flash cookies	**T6**. Canvas fingerprinting	**T7**. HTML5 local storage	**L1**. Lack of awareness (2)	**L2**. Lack of control (3)	**L3**. Bounded knowledge (4)	**Observations**
AdBlock Plus		✓	✓		✓			✓	✓	✓	Protects against some of the analyzed privacy threats, but threatens the economic model of the Web
Ghostery		✓	✓		✓			✓	✓	✓	Offers additional transparency functionalities to users regarding third-party tracking
Google Sharing		✓	✓								Aimed at protecting users only from cookie tracking performed by Google
Brave		✓		✓ (5)	✓	✓		✓	✓	✓	Based on the paradigm of a more user-driven ad platform; offers transparency and a great level of control to users
Subscribe2Web		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Posits a future ad-free Web, but it is unknown if this commercial solution necessarily implies stopping tracking users
Google Contributor		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Proposes an alternative economic model for the Web, but it is not clear if users will support content creators economically
Network-level ad blocking		✓	✓		✓	✓	✓	✓	✓		At network level, Internet providers are capable of offering similar blocking services than those of ad blockers

Table 3.10: Online advertising privacy threats and commercial solutions addressing them. (1) Since this threat derives directly from interactions explicitly triggered by users, protecting from it is a challenging task; (2) most solutions only show the number of trackers detected/blocked; (3) control is commonly enforced by given users blocking capabilities; (4) default configurations simplify their use but at the expense of privacy; (5) fingerprinting protection is currently in beta version.

a consequence of such ubiquity, online advertising has triggered the creation of a massive transport channel whose intermediary components have access to billions of users and, in particular, to their data. Even though gigabytes of aggregated user data support more targeted advertising campaigns, the inherent lack of transparency of online advertising entails serious risks to user privacy. In this chapter, by breaking down the instances of online advertising platforms and their corresponding capabilities (regarding user data), we have outlined an attacker model to describe the potential hazards to user privacy. We have emphasized the variety of information subject to be collected, the large number of intermediaries involved, their advanced and intrusive tracking capabilities, and the impact of advertising practices on privacy.

Unlike what happens with other online privacy threats, there is little users can do to completely prevent risks coupled with online advertising. Nevertheless, several solutions are offered to help to protect the privacy of users within such an opaque ecosystem. Accordingly, we have offered a wide range of mechanisms and we classified them into local solutions (browser and third-party based) and proposals based on new ad serving paradigms. On the one hand, some of the local solutions are very popular and their blocking approaches are already negatively impacting the economic model of online advertising. On the other hand, new advertising models are arising to offer native privacy and a stronger role for the user, while still proposing radical variations of the current advertising logic.

In addition, we have elaborated on the pros and cons of some of the aforementioned protection mechanisms with regard to the threats they try to alleviate within online advertising platforms. In such analysis, we also have outlined dynamic and smarter approaches proposed to avoid radical blocking mechanisms. Yet, based on the proposals analyzed, we have found it very hard to provide more privacy in the advertising ecosystem without significantly modifying the ad delivery model to give users more control and to reduce the financial dependence of Internet content on advertising.

Chapter 4

Privacy by Regulating the Distribution of Personal Data

4.1 Introduction

The growing access of people to information and communication technologies is contributing to reach the so-called "big data era", where the pervasiveness of data is a major input for increasingly personalized and automated online services. Among such services, online advertising aims at selecting and directing ads to the right potential customers (personalization) at the right time (real-time), built on multiple parameters, while users browse the Web [2–4]. Many details of this operation are given in Sec. 2.3.

This targeted advertising offers crucial benefits to several agents on the Internet. To start, users receive ads tailored to their interests and no longer static ads unrelated to their preferences; consequently, behavioral targeting are generating greater revenues than those of untargeted ads [26]. Furthermore, as previously described in Sec. 2.3, web sites (publishers) have access to an entire ecosystem to fund their operation through the money paid by demand side platforms (DSPs), which are advertising agencies acting in representation of advertisers[(a)]. Also, selling entities are given the

[(a)] When referring to the online advertising model in general, the terms 'buyer', DSP and 'advertiser' may be used interchangeably.

opportunity to promote their products over a ubiquitous structure with global reach. The upshot is that most of the content users consume online is supported by ad revenue.

We introduced in Sec. 2.4 one of the key enabling technologies that makes online advertising so profitable: real-time bidding, which enables advertisers to compete in real-time auctions to show their ads [25]. It is implemented by a management entity called ad exchange. Accordingly, when a user visits a website, her impression is sold to the advertiser (or corresponding DSP) that bids higher, in a matter of milliseconds. Moreover, DSPs are sent bid request messages containing user information (tracking data) to help them tailor ads to the user's preferences and decide the bidding strategy. In this way, RTB's aim is twofold: offering users a personalized experience through targeted ads and, thus, maximizing the profits of the whole advertising ecosystem. Whereas the operation of RTB behind the scenes is pretty opaque and complex for users [134], it is quite transparent for the actors of the advertising ecosystem. For example, ad exchanges provide DSPs with powerful management interfaces that offer very detailed information about the market and even enable buyers to set up their advertising strategies (e.g., by defining a targeting market). Certainly, a lot of benefits arise *for* the advertising ecosystem from the optimization capability offered by RTB in terms of automation, personalization, profit, and transparency.

Yet, despite its proven usefulness, the practices inherent to online advertising and RTB may pose serious privacy risks (see Sec. 3.2 for details) for users [17]. Most of these risks derive from the potential misuse of the user data flowing through the advertising ecosystem. To start, vast user data is mined at very high rates to implement real-time personalization; hence, truly detailed profiles are built about millions of people so fast and uncontrollably [74] that privacy protection is discouraged. Additionally, ad distribution mechanisms based on auctioning user impressions might lead to characterize users as more relevant (or more valuable economically) than others, depending on their profiles [14]; such a differentiation may entail social sorting or discrimination [15], thus an even less private environment. Finally, online advertising builds on interactions among myriads of intermediary ad companies that collect, use

and share user data, significantly increasing the risk of data misuse. Ironically, users have no control over how their data is managed in this context.

RTB builds on sharing user data with DSPs to encourage competition and ad personalization, but the unregulated distribution of such data may give rise to concerns. With the aim of helping DSPs decide whether to bid or not for a user impression, an ad exchange distributes to them personal information of the user whose impression is being auctioned (e.g., the URL being visited, the location of the user, or even a label categorizing the user). Thus, not only does the winner DSP receives this input, but also the rest of participating DSPs. This means that there could be agencies maliciously collecting data without even paying for it. We illustrate this risk in Sec. 4.2 where we unveil that a given DSP would have paid nothing for at least 55% of the users it tracked in a period of three months. This uncontrolled distribution of user data prompts a non-negligible privacy concern since an increasing number of advertising agencies are relying on RTB to daily reach billions of potential web customers [135]. Although the distribution of personal data among a group of DSPs cannot be entirely stopped without changing the current advertising business model, we report that the potential abuse of these agencies can be tackled with minimum tuning of said data distribution model.

Our proposal in this chapter builds on regulating the distribution of personal data from the ad exchange to DSPs when a user impression is auctioned. Such regulation consists essentially in limiting the number of DSPs invited to bid, that is, lowering the entities to which user data is leaked and, consequently, getting a more private environment. Accordingly, DSPs or similar intermediaries showing a dishonest behavior (e.g., never winning auctions) will be banned from participating in future auctions, which may entail correcting such harmful behaviors. At the same time, our approach strives to maximize the revenue of the ad exchange, looking for a balance with a given privacy level. The upshot is that some privacy can be reached without affecting the business model of the online advertising ecosystem, by slightly modifying the distribution of personal data among intermediary entities such as DSPs. The resulting adjusting effect on the behavior of these entities is relevant since privacy

concerns in general do not directly derive from the act of sharing data itself, but from the *inappropriate* sharing of user information [13].

Unlike our approach, other proposals have addressed this privacy issue through more radical strategies. Research proposals have concentrated on sophisticated mechanisms to anonymize or block the information leaked to third-parties while trying to remain compatible with the current ecosystem, but still requiring important modifications to its architecture and anyhow affecting its economy. On the other hand, commercial solutions have primarily focused on blocking tracking mechanisms at the cost of seriously damaging the Internet business model. However, as concluded in [17], it seems very hard to provide more privacy in the online advertising ecosystem without somehow modifying the ad delivery model.

4.1.1 Main requirements of the system

In this subsection we anticipate the main requirements around which our proposal in this chapter revolves, in order to guide the approaches we present in next sections.

- *Simple implementation.* To encourage its implementation in practice, we promote a solution that does not require modifying significantly the architecture of the online advertising ecosystem. This would imply a low adoption cost, unlike other academic approaches that propose rebuilding the current model to protect privacy.

- *Constructive technique.* Looking for an alternative to the arms race started by ad blockers, which is threatening the economy of the Web, we require a mechanism aimed to balance the tradeoff between user privacy and advertising utility (commonly in terms of money). This would limit the level of attainable privacy, but would open the door to further tangible mechanisms to address privacy concerns in this opaque environment.

- *Self-regulatory.* In line with a constructive approach, we uphold a system that allows misbehaving entities to correct their practices against privacy, under the

penalty of dynamic punishment. Namely, we require a solution that promotes appropriate practices towards user data to relieve privacy concerns.

Interestingly, the compliance with these three main requirements when designing our system will derive in additional aspects that may go in favor of user privacy.

The work presented in this chapter was published in [16].

Chapter outline

In this chapter, we illustrate the potential misuse of RTB with real data from a publicly available data set. We analyze the data of more than 64 millions of ad-auctions and interactions between a DSP and an ad exchange, to quantify the extent to which a DSP may collect user tracking data without paying for them. Namely, we performed a study reporting quantitative evidences on the misuse of RTB.

Since no preventive mechanism is currently put in place by Google's DoubleClick and AppNexus (the most relevant RTB systems), we hypothesize that such tracking and profiling practices may be rather common. To address this state of affairs, we design a system that aims to regulate the distribution of user data to third parties during the auctions for ad-impressions, i.e., to whom send the requests for each ad-space bidding.

The proposed solution is designed to strike a *balance* between *the average number of DSPs invited to bid* and the *revenue of the ad exchange* holding the auctions. Limiting the number of DSPs receiving user profiles naturally offer better privacy protection, especially since potential dishonest DSPs will hardly receive user sensitive information under such context. As a consequence, an ad exchange might be motivated to *suppress* the *bid requests* to abusing DSPs, but this would have an impact on its revenue. We formulate the problem of choosing a bid-request distribution as a multi-objective optimization problem that takes into account both aspects, i.e., the number of DSPs invited to bid and RTB profits.

We measure the extent to which user data is disseminated as the average number of DSPs receiving tracking data. Accordingly, for a desired data distribution strategy, our solution recommends, probabilistically and in real time, to which DSPs the ad

exchange should send a bid request for any given ad impression, in order to maximize the instant revenue. Evidently, with the aim of preventing abuses and thus supporting privacy, the fewer DSPs receive personal data the better. Experimental results show that our system seems to be able to tackle misbehaving DSPs.

The remainder of this chapter is organized as follows. Section 4.2 analyses the potential abuses and privacy risks we face in this context. Section 4.3 presents the theoretical analysis of our regulating approach. In Sec. 4.4, we evaluate our technique. Section 4.5 includes a relevant discussion about important topics of our approach and some general incentives to adopt it. Finally, conclusions are drawn in Sec. 4.6.

4.2 Privacy risks of Data Aggregation in Online Advertising

This section examines in depth the potential abuse and privacy risk object of this chapter. We emphasize that these issues derive from the capability of DSPs to track and profile users almost effortlessly and at very low cost. More specifically, user privacy in RTB systems is at risk as a result of: (1) user information is shared with third parties by default; (2) this information is not only delivered to the winner of an auction but also to other entities, and (3) there is an apparent lack of control over the abuse of potential malicious listeners.

Some guidelines are stated by the ad exchange (e.g., Google DoubleClick) regarding the use of auction data. Yet there are not known mechanisms to control such abuse from certain DSPs. Next, to illustrate the aforementioned privacy risk, we analyze a publicly available data set containing bid information of a Chinese DSP.

4.2.1 RTB: the auction technology behind online advertising

When a user visits a Web site with an ad space served through RTB [25], an HTTP request is submitted to the ad exchange, which subsequently sends "bid requests" to potential participants. We note that the number and type of participants involved

may vary on a per-auction basis, at the ad exchange's discretion. Within the bid request, the ad exchange generally includes the following data: the URL of the page being visited by the user; the topic category of the page; the user's IP address or parts of it; and other information related to their Web browser [12, 14, 35]. Accompanying this information, Google's ad exchange incorporates a bidder specific user ID, which implies that different bidders are given different IDs for a same user. Other RTB-based ad exchanges, alternatively, include their own user's cookies.

Upon receiving the bid request, the bidder may identify the user within its own database through the cookie or identifier. This is provided that the cookie-matching protocol has been executed previously for this user. Thanks to such cookie or identifier, the bidder can track them across those Web pages in which it is invited to bid [44]. From those tracked pages, the bidder can therefore build a profile, maybe complementing tracking and other personal data it may have about the user [41].

The bid price is then set on the basis of the bidder's targeting objectives, that is, whether it aims to target users visiting certain site categories, browsing from a given location, and/or having some specific profile. To evaluate if the ad-impression meets such objectives, the bidder relies on the aforementioned profile and the information included in the bid request. If interested, the bidder submits a price to the ad exchange, which finally, in a last step, allows the winning bidder to deliver the ad to the user. The winning bidder is evidently the highest bidder, but the price paid is the second-highest bid in the auction [136]; these so called second-price auctions look after preventing underbidding and overbidding. It is worth stressing that all this process of gathering user data, ad bidding and delivering is conducted in just tens of milliseconds.

4.2.2 Bid requests: the tokens leaking personal data

As explained in Sec. 4.2.1, a bid request not only serves to invite DSPs to participate in the auction of a user's impression. A bid request includes a variety of user data in order to provide DSPs with the necessary feedback to decide whether to bid or not for said impression. Then, interested DSPs send their bids to the ad exchange in order

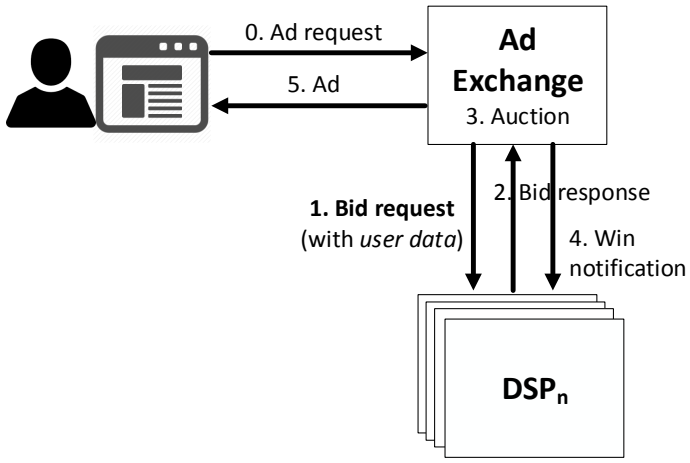


Figure 4.1: Interactions among an ad exchange and associated DSPs.

for an auction to be held. Evidently, the success of personalized advertising tightly depends on the granularity and volume of the information shared with DSPs. Sadly, user privacy decreases to the same extent that personalization improves. As an approach to evidence this privacy risk, we here portray the critical information available about the user and included in bid requests. Figure 4.1 depicts the aforementioned interactions among an ad exchange and DSPs. Considering that these interactions are carried out for every single user impression, it illustrates the wealth of personal information flowing to potential participants in the bidding process.

Dozens of fields and subfields carry information concerning the context in which a user impression is held [137]. As described previously, users play a leading role in this context. Thus, much of the information carried to fuel the RTB process characterizes them and, particularly, their behavior. First, a bid request may include a *user's ID* that DSPs may use to individuate them and match previously acquired information with the data included in bid requests. A user ID may be a string that unambiguously identify a user in a given system but not in real life, e.g., within the ad exchange's domain. Furthermore, the user's *IP address* (or part of it) is included in bid requests mostly to help DSPs infer *location information* to execute geographically targeted campaigns. IP addresses can also be used as user identifiers, especially now that IPv6 is providing an almost unlimited addressing space.

Additionally, device and web browser *fingerprint data* may be contained in a bid request, as powerful attributes to better identify users. A fingerprint is a set of attribute values that characterizes an entity to the point that could individuate it unequivocally. For example, a fingerprint of a network device might be composed by its operating system's name, its version, the list of applications installed and the list of open ports of the device.

Information about the users' *online behavior* may also be included in the bid requests sent to DSPs, e.g., in the form of a list of (user profile) *tags* or categories. These categories reveal the preferences of the user whose impression is auctioned, thus are crucial for DSPs when deciding whether to bid or not. Similar tags depicting the content of the website visited by the user might also be delivered to DSPs along with its *URL* or *domain*. Finally, a *time stamp* indicating the date and time of the user's visit, and a reference *bidding price* to inform the minimum value to bid may be provided by bid requests.

Several privacy risks may derive from this personal information, especially when distributed among several intermediate entities such as DSPs, in a position to aggregate and process said information. To start, although user IDs do not identify a user in real life, a combination of the remaining attributes may inequivocally individuate a user (a few demographic attributes have such an identification power [138]). Users' location information could lead an attacker to learn moving patterns of users to then reveal even further details about their daily activities [139]. Device and web browser fingerprints may complement this attack by enabling cross device tracking [140]. Not only could users and their activities be geographically tracked using data in bid requests, but also their preferences are learned and may reveal sensitive information[141]. In fact, pricing information is already a critical aspect that directly discloses the relative importance of a user. Table 4.1 maps some of these information items to the potential privacy risks derived from their open distribution and aggregation.

Table 4.1: Information, items carried in bid requests, here matched to the potential privacy risks derived from their open distribution and aggregation.

	Identification	Learning of moving patterns	Cross device tracking	Micro targeting	Habits tracking	Outlier detection
user ID	x					
IP address		x				
user location		x	x		x	
device fingerprint	x		x		x	
web browser fingerprint	x		x		x	
time stamp		x				
user languages	x			x		
user labels				x		x
URL					x	
content labels				x		
minimum bid price			x			x

4.2.3 The iPinYou data set

To illustrate the potential misuse of RTB systems and its real impact on user privacy, we analyze a data set that includes bid information released by a well-known Chinese DSP called iPinYou.

The iPinYou data set [142] contains logs of the ad auctions where this DSP has participated. These logs basically carry three types of information for each auction: (1) user data sent by the ad exchange to the DSP in order for the latter to prepare a bid response, (2) the price paid when it wins the auction, and (3) information on whether the user made a click or a conversion as a response from the ad displayed. User data include some of the parameters described in Sec. 4.2.2, e.g., an ID of the user that generated the auction, a timestamp, their browser fingerprint (user-agent), their IP address (its first 3 bytes), their location (region and city), the domain and URL visited, and some user tags representing the categories of interest of the user. Additional information involves the ad exchange that held an auction and the price paid by a DSP (not necessarily iPinYou) to won it. The values of some of these attributes, e.g., IP address, URL and domain visited, are anonymized to preserve the privacy of users. It is important to note that this data set contains information related to the bids in which iPinYou participated, excluding the auctions where iPinYou decided not to bid.

Table 4.2: Parameters describing the iPinYou data set. This includes, e.g., the number of bid log records, unique users involved, or the number of Chinese cities reached.

Bid log records	64.746.749
Logs of won bid requests	19.495.976
Unique users	21.264.865
Data attributes in bid logs	24
Ad exchanges	3
Regions	35
Cities	362
User profile tags	44

This data set was released in 2013 for an open contest on RTB ad pricing. For the purpose of our analysis, we use the version processed by Zhang et al in [142]. We aggregated the data from seasons 2 and 3 of the competition (data from season 1 has different fields than the rest) and we examined the data of almost 65 million bid requests sent to this DSP. We found that these bid requests belong to about 21 million unique users. In Table 4.2 we summarize the most relevant figures of the data set at hand. In order to facilitate the processing of this data, we used a sample of bid requests corresponding to the users having 70 or more log records in the whole data set, yielding almost 6 thousand users with more than 8 thousands log records.

4.2.4 Privacy risks and abusing context

User privacy risk starts from the capability of an ad exchange to *identify* the user whose impression is being auctioned. The user ID attribute included in bid requests and thus sent to DSPs unequivocally identifies a user within that context. In fact, if this identifier is already known by a DSP, they could match even more information about the user. In addition, recall, e.g., that a few combined demographic attributes may be very identifiable. Consequently, other attributes such as the user's IP address and the device fingerprint [46] might make this risk worse. Namely, although not real-life identifiers, user IDs, when combined with other bid request fields of information, might significantly facilitate the work of a privacy adversary in its bid for individuating a victim.

Public IP addresses could by themselves be very identifiable, too. For this reason, only the first three octets are commonly revealed in bid requests, but it is still evident

when the address changes. The uniform change of a user's IP address through the day, if a user is tracked across different geographical areas, might unveil movement patterns, which is sensitive information with regard to user privacy [143]. With respect to this, within the iPinYou data set, we found that a great portion of users (about thirty percent) were associated with three or more different IP addresses.

In addition to IP addresses, other attributes with rare attribute values may help adversaries single out users in real life, even more when analyzed in combination with other attributes. For example, people using Linux operating systems and non standard web browsers (e.g., Opera) could excel so much to become easily identifiable outliers. To have an idea of this, in the iPinYou data set, we found only 206 bid requests (out of millions) including user information coupled with the combination of Linux operating system and the Opera browser.

This process of associating a user's unique identity with their interactions enables *tracking*. Then, working in real time, tracking allows advertising entities to "recognize" users during their impressions and ultimately display a personalized ad. However, tracking also enables these entities to join personal information to build individual user profiles (*profiling*). As in other personalization contexts, such tracking and profiling capabilities are supported by the processing of personal information. Nevertheless, within advertising platforms, personal information flows freely, constantly, and abundantly from the ad exchange to DSPs. Thus, a sort of *oversending of personal data* to third parties might be supporting misuse and worsening privacy risks.

The last statement implies that DSPs essentially do not pay for the user information they receive in bid requests, but for the auctions they win on behalf of advertisers. In practice, upon the reception of bid requests (invitations to participate in auctions), a DSP pays just for the auctions it wins, while it receives user data in the rest of bid requests "for free". Clearly, DSPs may take advantage of the ad exchange's tracking resources at a very low cost. This fact is evidenced in Fig. 4.2, where we depict the amount of users whose information has been paid by the iPinYou DSP. To illustrate the amount of information potentially collected for free, we can see in this figure that, *for about 55% of the users, this DSP has not paid for any of their bids*. From this

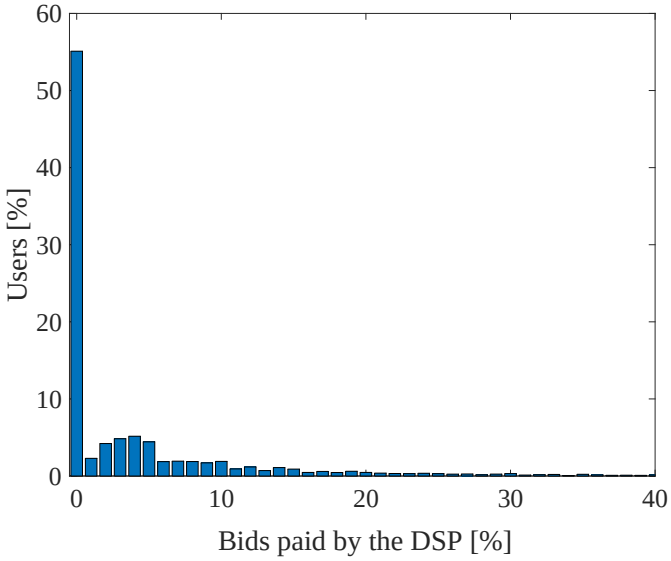


Figure 4.2: Percentage of users whose information has been paid by the iPinYou DSP. For about 55% of the users, none of the bid requests triggered from their impressions were paid by the DSP, i.e., the DSP did not pay anything for the auctions held for 55% of the users.

we can infer the potential abuse of a third party and the exacerbated risk for the privacy of users if multiple DSPs exhibit a behavior not oriented to participate in auctions, but to take advantage of the large amount of user data distributed by an ad exchange. We would like to stress, however, that this percentage of users tracked for free might be just a lower bound: the released data set does not include the auctions where iPinYou did not bid, but from which it could have received numerous user data costing nothing.

In an attempt to prevent this abuse, ad exchanges clearly prohibit DSPs to use the information in bid requests when a corresponding auction has not been won [144]. It is also not allowed to use this information for applications other than the ones related to online advertising. However, enforcing such use is hard when the information has already been distributed to third parties; and when, due to an increasingly complex advertising ecosystem, more and more entities are included to outsource specific functions in the demand side (e.g., trading desks).

Data aggregation performed by intermediate entities brings another privacy jeopardy in online platforms. As explained in Sec. 4.2, not only ad exchanges, the core

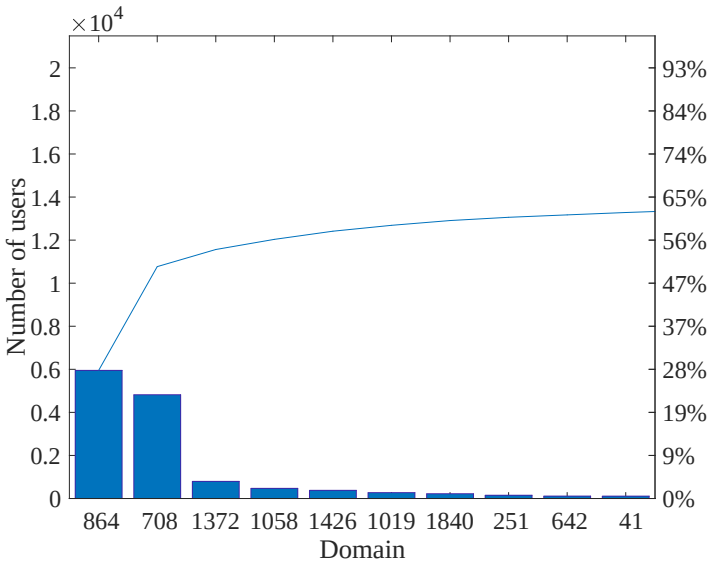


Figure 4.3: Amount of users tracked by the most popular domains in the iPinYou data set.

of ad distribution, but also DSPs and even publishers are in a position to *concentrate user data*. As expected, in the iPinYou data set, user tracking is concentrated in Google's DoubleClick ad exchange. Furthermore, we found that more than fifty percent of the users in the iPinYou data set would be tracked by only two publishers, probably related to the most popular websites in China (Fig. 4.3). In other words, having recognized at least 2063 publishers in this data set, less than 0.1% of them concentrates the tracking of more than 50% of the involved users. Powerful *tracking* capabilities are then held in very few hands.

Not only the easiness and openness of data collection is threatening the privacy of Internet users, but also the level of detail of the data. The *granularity* of the user data held by these entities has given rise to powerful capabilities of microtargeting. These capabilities have derived in tools to select audiences that may enable even advertisers to target groups of users with great precision [145]. In Fig. 4.4, we show an interface offered by a social network and a DSP to choose an audience for better ad targeting.

Finally, due to the pervasiveness of online advertising, it is not hard to comprehend its wide reach in the population. The idea that the advertising ecosystem might be collecting information related to large masses of people is reflected in the iPinYou data set. More precisely, based on the user ID and region attributes of the

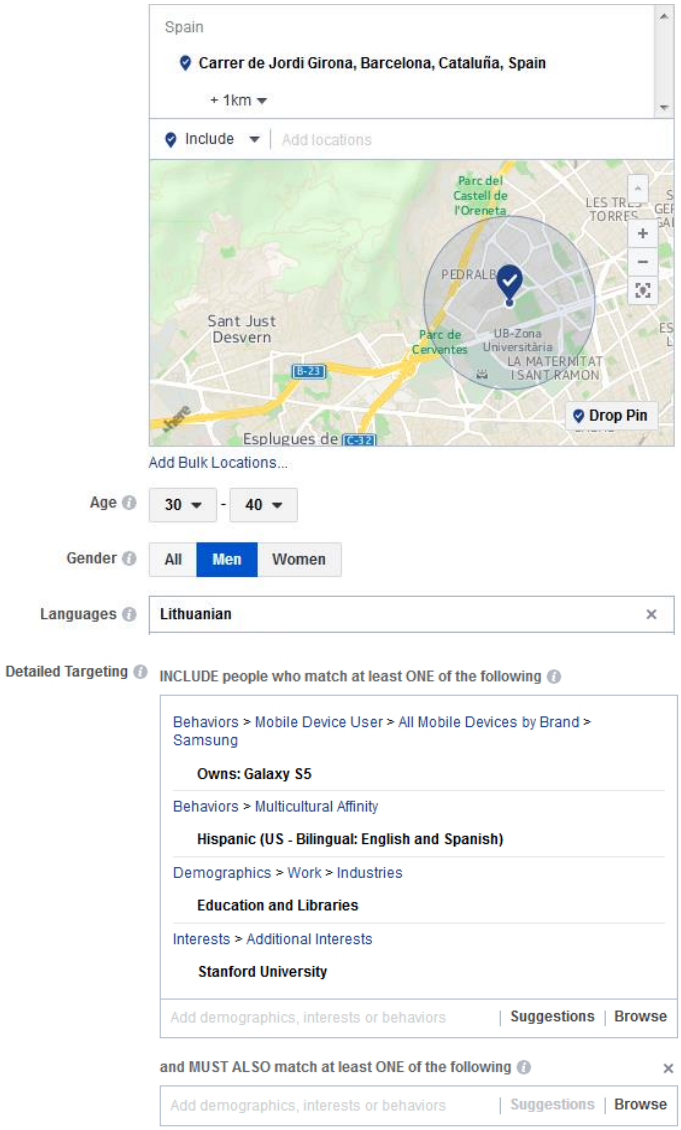


Figure 4.4: Interface for advertisers to select an audience for a campaign in Facebook. Its very granular options allow a great power to microtarget users.

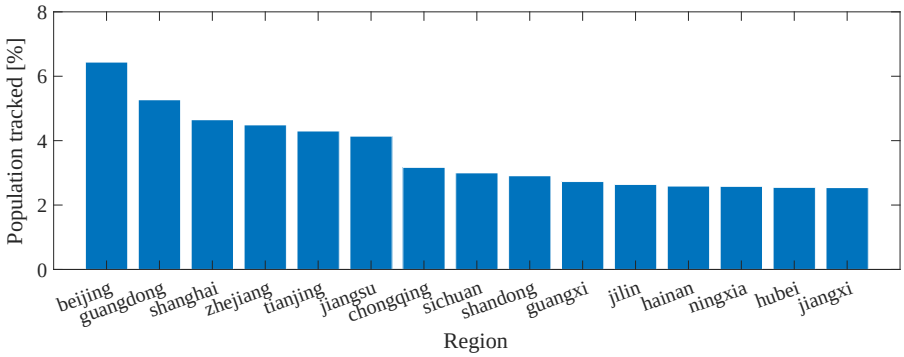


Figure 4.5: Population tracked by online advertising entities along different regions of China, as observed in the iPinYou data set.

records from this data set, we observed that large portions of the population of important Chinese regions would have been tracked. For example, this DSP (iPinYou) would have information of more than 5% of the population in regions such as Beijing, Guangdong, and Shanghai (see Fig. 4.5). Considering that, in this case, most of the user data must be "ceded" as input to DSPs for their bidding decisions, gathering such bulks of information seems a very good deal for them. However, this is not good news for the privacy of users, who are probably being observed en masse.

4.3 Controlled Distribution of Bid Requests

Along this section, we propose a system that aims to reduce the oversending of personal data to DSPs, thus ultimately providing some privacy to users in RTB systems. This is done by regulating the distribution of bid requests among intermediary entities such as DSPs or trading desks. Conceptually, this objective could be reached, to a certain extent, by reducing the number of DSPs to which bid requests are sent, thus lowering the instances where user data is aggregated. Naturally, from a practical perspective, our solution is conceived to be implemented within the ad exchange infrastructure since it is the entity in charge of sending bid requests to DSPs when a potential ad impression arises. The proposed system determines, in real time and adaptively, the specific participants of a given ad-space auction, at the cost of some processing overhead at the ad exchange and a potential reduction in revenue incurred by a smaller number of participants in auctions. Being revenue the *raison d'être* of ad

exchanges, a trade-off will arise with data distribution control, but with an adequate balance we show that reasonable guarantees can be provided while keeping relatively high profits.

Unlike many of the privacy techniques proposed in the literature for online advertising, a change in the distribution model of bid requests does not entail an important modification of the advertising ecosystem.

4.3.1 Adversary model

Our technique builds upon the principle of a selective distribution of bid-request information (containing user sensitive data) among potentially interested DSPs. Consistently with this principle, we assume an adversary model in which the bid requests sent by an ad exchange are passively observed and maliciously aggregated by a group of intermediary entities.

We must stress that this adversary model assumes that privacy risk comes from the exploitation of user profiles built from the *aggregation* of user data. Namely, the user data in a single bid request would not entail a significant privacy risk since by itself reveals only a snapshot of the preferences, behavior and demographics of a user at a certain point in time. However, the more user data is aggregated the richer are the resulting profiles, and the higher is the corresponding privacy risk.

As argued in Sec. 4.2, RTB-based ecosystems still provide fertile ground for privacy abuse. One of the reasons is the relative ease with which user data can be collected by intermediate and authorized entities such as DSPs and other smaller subsidiary entities (e.g. trading desks). Especially the latter, sometimes being really small companies, are becoming capable of tracking users at a very low cost (or none) and without deploying an important infrastructure. Thus, a privacy gap arises when they are given easy access to an ever-growing universe of aggregated personal data. We propose a system to bridge this gap by penalizing said kind of tracking when it violates the norms established by ad exchanges.

As a note, abusing of such tracking is against the terms of use of the main ad exchanges. These terms forbid DSPs taking advantage of data for which they have

Table 4.3: Behaviors of DSPs, with regard to their participation in bid auctions, that may go against ad exchanges policies and also in favor of the violation of users' privacy

Behavior	Description
silent	A DSP not participating in auctions, and thus not answering bid requests, may be misusing the RTB infrastructure by collecting and exploiting the user data carried in theses messages. Ad exchanges recognize said risk when forbidding DSPs using the data for which they have not paid in their policies [144]. Although this also gives rise to privacy concerns, no further control is made.
loser	A DSP that looses too many auctions is also suspicious of abusing the RTB infrastructure against privacy. Bidding to loose is possible since bid requests sent to DSPs include information about the minimum price of the user impression auctioned. Just by bidding below the minimum would enable DSPs or related entities to receive user data for free.
stingy	A DSP bidding too low might be looking for receiving user data when no pricing information is received in bid requests, thus trying to inappropriately exploit the RTB logic in detriment of privacy.

not paid. For example, according to Google DoubleClick Ad Exchange (AdX) Buyer Program Guidelines [144], some of the policies that buyers must adhere to are listed next:

- Buyers and any third party to which they provide access to the ad exchange must adhere to the policies.
- The inventory purchased cannot be sold to another sales channel.
- Bid data cannot be used for purposes different from buying on the ad exchange.
- Unless a buyer wins a given impression, it must not use bid data for that impression to create user lists or profile users.

In brief, neither DSPs nor outsourced entities such as trading desks are allowed to exploit bid data coming from an ad exchange, unless they have paid for such data after winning a given auction. Some of the behaviors that might go against ad exchange's policies are described in Table 4.3.

4.3.2 Bid request distribution model

As noted in Sec. 4.2.1, the visit of a user to a website that holds an ad space generates a so-called ad impression. Then, an ad exchange auctions said impression among all the available DSPs. To support the bidding decision of DSPs, the ad exchange distributes among them bid requests containing some user data.

We propose reducing the number of DSPs to which a bid request is sent, in order to penalize misbehaving DSPs and to promote privacy. To model the distribution of bid requests, we rely on the Bernoulli distribution that characterizes a discrete probability distribution of a random variable whose value is *true* with probability p and *false* with probability $1-p$. This is the same behavior of the outcome of sending bid requests to DSPs; they will receive requests if behaving well or will not receive bid requests (penalized) if being dishonest. Accordingly, being d the number of DSPs available in a given moment, we model the distribution of bid requests among them as the execution of d Bernoulli trials (or experiments).

These trials can be represented as d independent, identically distributed Bernoulli random variables (r.v.'s) $X_1, \ldots, X_d$, each of which characterizes an experiment with a boolean-valued outcome and a success probability p_i, with $0 \leq p_i \leq 1$. Therefore, when auctioning a user impression, the ad exchange shall send a bid request to DSP_i with probability p_i and shall not do it with probability $1-p_i$. A given ad exchange's distribution strategy will be defined as the tuple $p = (p_1, \ldots, p_d)$ of the probabilities of sending a bid request to each of the d available DSPs. In Fig. 4.6, we depict this distribution model for a given user impression.

As introduced previously, to control misbehaving DSPs, we propose bounding the number of DSPs that receive a bid request from the ad exchange. Intuitively, the less the number of receiving DSPs, the higher the level of user privacy. To do it, we introduce a *data distribution control parameter* defined as the average number of DSPs that will receive a bid request, α, with $0 \leq \alpha \leq d$. Namely, in our system, the number of recipient DSPs is bounded to the value of α. Clearly, the number of invited DSPs, being a sum of independent Bernoulli trials, follows a Poisson binomial distribution with mean $\sum_i p_i$. Consequently, our measure of privacy, the average number of participating DSPs, can be computed straightforwardly as $\alpha = \sum_i p_i$.

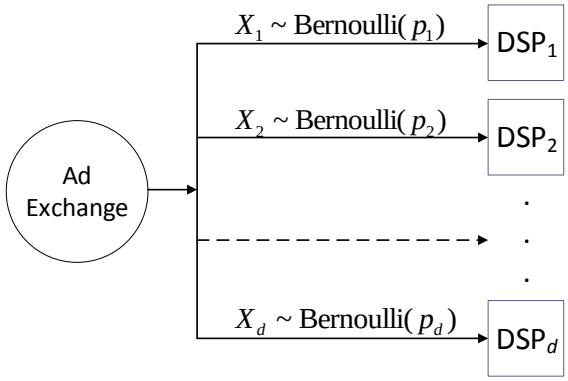


Figure 4.6: A depiction of the bid request distribution model we propose for the ad exchange. We model said distribution as the random draw of d Bernoulli trials (represented with d Bernoulli r.v.'s: $X_1, \ldots, X_d$), being d the number of DSPs available. Each r.v. characterizes an experiment with a boolean-valued outcome and a success probability p_i.

4.3.3 A system to balance the number of DSPs invited and ad revenue

In Sec. 4.3.1 we described the adversary model we tackle in this chapter. In particular, we mentioned that DSPs might go against the policies of the corresponding ad exchange by exploiting a uncontrolled distribution of bid requests. Nevertheless, implementing these policies is by no means a simple task because ad exchanges have no control over the internal dynamics of buyers' data infrastructures. In any case, they do have the capability to regulate how bid requests are distributed to buyers. Then, it is by shaping such distribution of user data, according to the behavior of DSPs, that we propose to bound the amount of information (bid requests) sent to DSPs, with the ultimate aim of enhancing privacy.

Intuitively, a distribution strategy that restricts the recipients of bid requests will reduce the revenue of an ad exchange. Accordingly, we define a metric of said revenue, in a given auction, as the product of three variables ω_i, μ_i and p_i, for $i = 1, \ldots, d$. Note that maximizing this measure of revenue would imply maximizing the real revenue, according to the distribution model proposed in this chapter. Both ω_i and μ_i are system parameters taking values in $\mathbb{R}$ and could be interpreted as reputation metrics of a DSP i. A DSP that behaves according to the ad exchange's rules will generate

a reasonable revenue and thus will have a better reputation than other DSPs that break the rules. For each DSP i, we define the *winning rate* ω_i as the rate of won bids with respect to the number of bid requests received up to a given instant. Weighting by winning rate enables our model to discourage a potential misuse of the bid request distribution model in online advertising. A DSP that almost always loses is probably just "listening" for user data to tamper with their privacy, thus deceitfully exploiting the online advertising ecosystem. In our proposal, the economical contribution of a DSP winning only a few auctions, even bidding high, will be weighted by its poor wining rate in order to counteract its behavior against privacy.

In addition, we define μ_i as the average money spent by a DSP up to a given instant, that is, the amount of money paid for the won bids divided by the number of bid requests received (*average money spent*). Next, for the sake of simplicity, we refer to the product of ω_i and μ_i as r_i. For the sake of clarity, please refer to Table 4.4 to find the notation used in this analysis.

We denote by p the strategy of distributing bid requests where p_i, already defined in Sec. 4.3.2, could be seen as the percentage of traffic sent to DSP_i. Evidently, the higher the winning rate ω_i and the average money spent of a DSP_i, μ_i, the more likely it is to win an auction (thus having a higher "reputation"). Naturally interested in reaching the maximum possible revenue, an ad exchange will try to send a bid request to the DSPs with the highest product r_i. However, for DSPs with low r_i, i.e., showing bad behavior, in order not to completely eliminate their opportunity to participate in auctions, we will impose a *tolerance* parameter, i.e., a lower bound on p_i, denoted by $p_{\min} > 0$. Thus, with $p_{\min} \leq p_i$, we try to guarantee, for said DSPs, the chance to improve their behavior (reputation) in the future.

According to the justifications in Sec. 4.3.2, in our approach we use the parameter α to bound the number of DSPs invited to bid (invitation rate) and that will receive information from the ad exchange. Put another way, α could also be interpreted as a measure of the suppression of bid requests to DSPs. Consistently with this bound, we define a revenue-invitation rate function

Table 4.4: Description of the main variables used in our notation.

Symbol	Description
d	Number of DSPs available in our scenario
p	Tuple representing the ad exchange's distribution strategy. Its elements are the probabilities of sending a bid request to each DSP
α	Average number of DSPs that will receive a bid request, i.e., the average number of DSPs to be invited to the auctions
ω_i	Rate of won bids with respect to the number of bid requests received by a DSP i up to a given instant
μ_i	Average money spent by the i-th DSP up to a given instant
r_i	The product of $\omega_i \mu_i$
$p_{\min}$	Lower bound on p_i that guarantees an opportunity to participate in auctions for all DSPs
$\mathcal{R}(\alpha)$	Function modeling the revenue of an ad exchange as a function of the privacy parameter α

$$\mathcal{R}(\alpha) = \max_{\substack{p \\ p_{\min} \leq p_i \leq 1 \\ \sum_{i=1}^{d} p_i = \alpha}} \sum p_i \omega_i \mu_i, \tag{4.1}$$

which characterizes the optimal trade-off between *revenue* $\mathcal{R}$ and the number of DSPs invited to bid α. From this expression, we aim at finding an optimal strategy of bid request distribution p^*, that satisfies an average participating DSPs α while maximizing the resulting ad exchange's revenue $\mathcal{R}$. Note that this expression establishes a strict restriction (it must be fulfilled) regarding the limit of DSPs that will receive invitations by the ad exchange (α), while its revenue is maximized in a best-effort sense. We would like to stress that the priority in our definition is meeting this bound, i.e., to prevent abuse and mitigate the privacy risk.

Although we propose modulating (or restricting) the distribution of bid requests to DSPs such that more privacy is provided to users, this does not necessarily imply that ad exchanges lose control to exploit user data. In fact, our approach would leave unchanged the internal logic within ad exchanges for the sake of simplicity and applicability; this includes how user data is collected and processed by ad exchanges. Our proposal focuses rather on the flow of user information from ad exchanges to DSPs, since unnecessary interactions threatening privacy may arise in such data sharing context.

Having presented the main parameters and indicators of our system, we summarize in the next list of steps the actions that the ad exchange must perform to integrate our approach to the auctioning system. Also Fig. 4.7 illustrates this integration and later on is used to describe the evaluation methodology of our bid request distribution strategy.

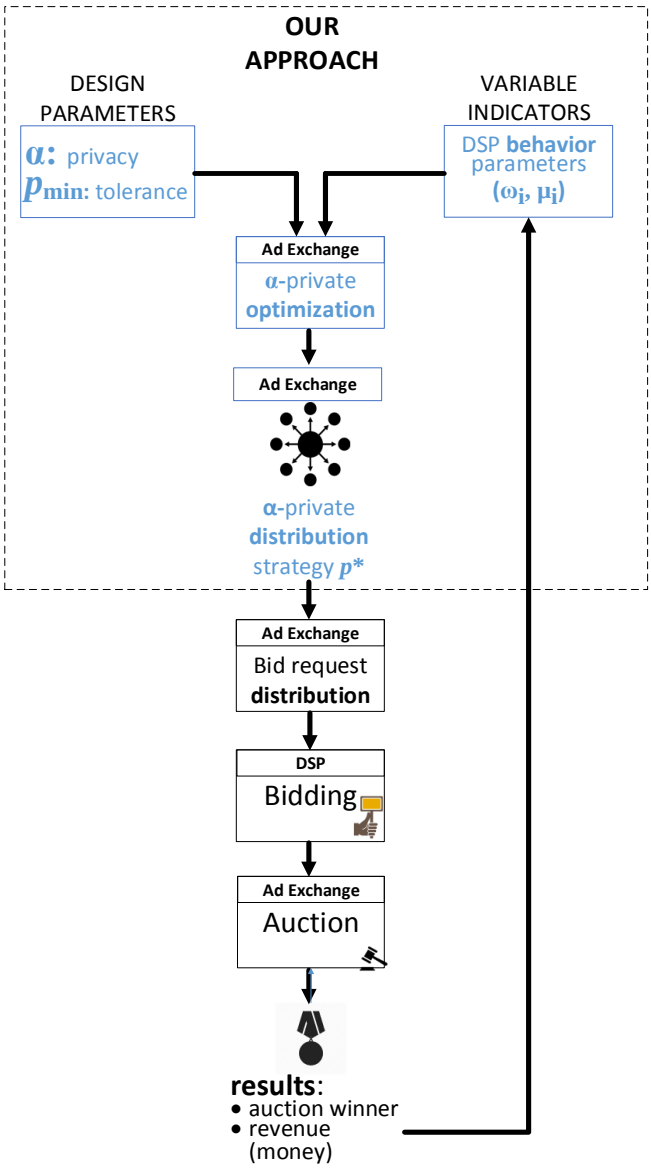


Figure 4.7: Methodology implemented to assess our bid request distribution strategy. This flowchart also illustrates how our system integrates to the ad exchange's auctioning system as described in Sec. 4.3.3 (the blocks in blue).

Step 1: Set the design parameters of the system: a bid request distribution (privacy) parameter α and a tolerance parameter $p_{\min}$.

Step 2: For each d DSPs, compute and update their variable behavior (reputation) indicators based on their win rate and money spent (ω_i, μ_i).

Step 3: Find an optimal distribution strategy of bid requests $p^* = (p_1^*, \ldots, p_d^*)$ that balance a measure of privacy with revenue.

Step 4: Send bid requests (invitations) only to the α DSPs showing the best behavior indicators.

Step 5: Receive bid responses and auction the user impression.

4.3.4 Optimal strategy for the distribution of bid requests

In this section, we analyze the revenue-invitation rate function (4.1) defined in Sec. 4.3.3, and present a closed-form solution, albeit piecewise, to the maximization problem. We suppose, without loss of generality, that

$$r_1 \geq \ldots \geq r_d. \tag{4.2}$$

Also, for $k = 1, \ldots d$, we define a sequence of thresholds α_k as $k(1 - p_{\min}) - d\, p_{\min}$.

Lemma 1. For any $k = 1, \ldots, d$ and any $\alpha \in [\alpha_{k-1}, \alpha_k]$, the solution to (4.1) is the distribution strategy

$$p_j^* = \begin{cases} 1 & , j = 1, \ldots, k-1 \\ \alpha - p_{\min}(d-k) - (k-1) & , j = k \\ p_{\min} & , j = k+1, \ldots, d \end{cases} \tag{4.3}$$

and the corresponding maximum revenue yields

$$\mathcal{R}^*(\alpha) \;=\; r_k\alpha \;-\; r_k p_{\min}(d \;-\; k) \;-\; r_k(k \;-\; 1) \;+\; \sum_{j=1}^{k-1} r_j \;+\; p_{\min} \sum_{j=k+1}^{d} r_j. \quad (4.4)$$

Proof. The existence and uniqueness of the solution is a consequence of the fact that we maximize a continuous function over a compact set.

From the assumption (4.2), it follows intuitively that for an $\alpha < 1$ the solution consists in sending a bid request to the first DSP, i.e., to the DSP having the maximum product $\omega_i\mu_i$. However, the condition $p \geqslant p_{\min}$ ensures the resource, α, must be distributed across all other DSPs, so that all participants can have a chance to receive a bid. The amount of α to be distributed is clearly $dp_{\min}$ and hence the remainder $\alpha' = \alpha - -dp_{\min}$ is the resource to be assigned among the d DSPs. Therefore, $p_{\min} \leqslant \frac{\alpha}{d} \leqslant 1$.

Following the same intuitive principle described above, we proceed to examine the distribution strategy of the remaining α'. Note that, below, all the expressions in terms of α' can be recast in terms of α' on account of $\alpha = \alpha' + p_{\min}$. For notational convenience, define $p' = p - p_{\min}$.

Case 1. $0 \leq \alpha' \leq 1 - p_{\min}$.
Observe that, in this case, any feasible $p' = (p'_1, \ldots, p'_d)$ satisfies

$$p'_1 r_1 + \ldots + p'_d r_d \leq (p'_1 + \ldots + p'_d) r_1 = \alpha' r_1,$$

which implies that the optimal distribution strategy consists in assigning the whole α' to the first DSP, that is $p^*_1 = \alpha'$ and $p^*_i = 0$ for $i \neq 1$. More compactly,

$$p^* = (\alpha - (d-1)p_{\min}, p_{\min}, \ldots, p_{\min}),$$

by virtue of $\sum p^*_i = \alpha = dp_{\min} + \alpha'$.

Case 2. $1 - p_{\min} < \alpha' \leq 2(1 - p_{\min})$.
This case follows in an exactly analogous manner as the previous case and leads to the optimal strategy

$$p^* = (1, \alpha - (d-2)p_{\min}, p_{\min}, \ldots, p_{\min}).$$

Case k**.** $(k-1)(1 - p_{\min}) < \alpha' \leq k(1 - p_{\min})$.

By generalizing our analysis for the k-th case, written in terms of α as $(k-1)(1-p_{\min})+d\,p_{\min} < \alpha \leq k(1-p_{\min})+d\,p_{\min}$, it is straightforward to check that the optimal distribution strategy is

$$p^* = (1, 1, \ldots, \alpha-(d-k)p_{\min}, p_{\min}, \ldots, p_{\min}).$$

Simple algebraic manipulation leads to expression given in the lemma. The derivation of the maximum revenue follows immediately from the optimal strategy as $\mathcal{R}^*(\alpha) = \sum_{j=1}^{d} p_j^* r_j$. □

The optimal bid request distribution strategy in Lemma 1 is interpreted as follows. Given the first condition of our problem (4.1), $\sum_{i=1}^{d} p_i = \alpha$, the average number of DSPs α to which requests will be sent, has to be distributed among the d available DSPs. According to (4.3) in Lemma 1, the first $k-1$ DSPs (the ones bidding more and winning more auctions) are by default sent a bid request; the probability of sending them the request is 1. The last $d-k$ DSPs (the ones bidding less and winning less auctions), however, are sent a bid request with a minimum probability $p_{\min}$ according to the first condition of our revenue-invitation rate function (4.1). Finally, the k-th DSP is sent a bid request with the remaining probability $\alpha - p_{\min}(d-k) - (k-1)$. This strategy can be easily explained as a resource allocation problem where α (the "resource to be distributed") is shared among DSPs according to their good behavior, with the aim of satisfying a given bound α.

Next, we proceed to analyze very briefly some properties of the revenue-invitation rate function (4.4). It is immediate to check the function is piece-wise linear with slopes $\omega_k \mu_k$. Given that this product will never be negative, neither will be the slope of $\mathcal{R}(\alpha)$ and, consequently, it is easy to see that $\mathcal{R}(\alpha)$ is nondecreasing. We cannot characterize $\mathcal{R}(\alpha)$ as increasing because there is the possibility that $\omega_k \mu_k$ is zero. Under the same reasoning, it is immediate to check the monotonicity of this function. Also, from Lemma 1, it is routine to verify the continuity of $\mathcal{R}$ on the interval $\alpha \in [1, d]$. To show the convexity of $\mathcal{R}(\alpha)$, note again that for each k and $\alpha \in (\alpha_{k-1}, \alpha_k]$, the optimal tradeoff function has slope r_k (or $w_k u_k$). From the labeling assumption (4.2), it follows immediately that $\mathcal{R}(\alpha)$ is defined by the decreasing sequence of positive slopes $r_1, \ldots, r_d$ (or $w_1 u_1, \ldots, w_d u_d$) and therefore is

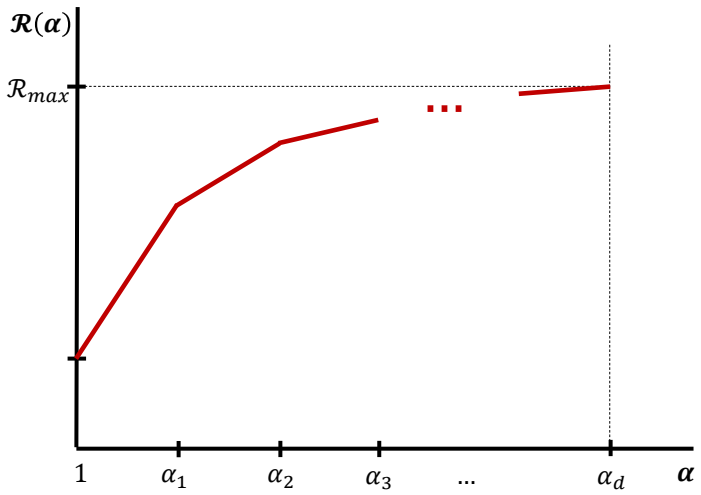


Figure 4.8: Conceptual plot of the revenue-invitation rate function. $\mathcal{R}(\alpha)$ is a nondecreasing function defined piecewise. From the labeling assumption (4.2), the slopes of $\mathcal{R}(\alpha)$ ($\omega_k \mu_k$) decrease as α grows.

concave. Fig. 4.8 conceptually illustrates these properties and the results of Lemma 1.

From the plot in Fig. 4.8 we can observe that the behavior of the DSPs (graphically depicted through the slopes r_k) determines that the losses in revenue of the ad exchange could be rather low. Namely, the higher the slope r_k (i.e., the better the behavior) of the first DSPs, the faster $\mathcal{R}(\alpha)$ approaches the ideal revenue $\mathcal{R}_{\max}$. This would entail a more controlled and potentially private bid request distribution (since fewer DSPs would be involved) while not significantly impacting the revenues of the advertising ecosystem. The notation used in this chapter is summarized in Table 4.4.

4.4 Experimental Analysis

Next, we empirically evaluate the solution proposed in Sec. 4.3. We describe our experimental methodology and outline the scenario simulated to reproduce the bidding process performed by an ad exchange and a set of DSPs. This allows us to investigate the effect of modifying the bid request distribution model with the aim to enhance user privacy. Our analysis also contemplates measuring said impact in the revenue of the ad exchange.

Since our proposal is to reduce the potential ad buyers to which user data flows, an impact is expected on the revenue obtained by the online advertising ecosystem from these bidders. In particular, given the importance of the advertising business model for the operation of the Internet, we need to show that our proposal does not significantly affect said business model. Accordingly, when applying our strategy, we expect a reduction on the revenue for the ad exchange. However, supported by the optimization approach described in Sec. 4.3.3, we also need to verify that this loss in income is acceptable in light of the benefits of a more privacy-respectful system[(b)]. Furthermore, we also verify if, as a result of our multicast solution, the misuse of RTB systems by some DSPs can be effectively addressed.

4.4.1 Experimental methodology

The proposed solution affects the interaction among an ad exchange and associated DSPs. Recall from Chapter 2 that DSPs act on behalf of advertisers and thus as bidders (buyers) in the auctions organized by an ad exchange. In order to invite DSPs to participate in a given auction and to provide them with the necessary feedback, the ad exchange distributes bid requests among them, including detailed data about the user whose impression (and corresponding ad space) shall be auctioned. This distribution of user data is adjusted in our approach with the objective of preventing dishonest behaviors of data collection and thus trying to preserve privacy.

To validate our mechanism, we configure an auctioning scenario that reproduces this behavior, through a Matlab simulation. In this scenario, considering a distribution control parameter α, an ad exchange enables a number of DSPs to participate in each auction, while optimizing its revenue. The main elements of this setup are depicted below.

In our experimental methodology, we simulate real-time auctions in which a variety of DSP types may bid. In order to deploy a more realistic setup, we consider three

[(b)] As commented in Sec. 4.1, a great portion of users use ad blockers for privacy reasons. If the proposed system meets the requirements of these privacy-aware users for some α, the loss in revenue due to our multicast strategy (instead of the current broadcast approach) may be more than compensated by the gains of these non-blocking users.

types of DSPs according to the more likely value of their bids: DSPs bidding high, low, and average. For each auction, the bids from every DSP are randomly sampled from a range of values reflecting these behaviors. For our experiments, such bids are generated following both uniform and Gaussian distributions.

After bids are generated probabilistically at every time instant, an ad exchange instance holds an auction and determines the winner DSP (the one with the highest bid). In line with our privacy proposal, for every auction, not all available DSPs are "invited", i.e., not all DSPs are sent bid requests, but a number of them, according to the parameter α. Thus, the corresponding activation of DSPs to participate in every auction is enabled by the optimized distribution strategy defined in (4.3). The strategy depends on two parameters specific to the historical operation (behavior) of each DSP (winning bid rate, and average money spent) and on the privacy parameter α defined by design. Consistently, said parameters of each DSP are calculated before an auction to be used as a kind of reputation metric that fuels the private bid request distribution strategy. Figure 4.7 depicts this methodology implemented through a simulation using Matlab R2017a.

After simulating one thousand auctions, we compute the total revenue of the ad exchange by summing all the money effectively spent by the bidders that won at every time instant.

To evaluate if our approach is feasible, we need to examine to which extent it may impact the ad exchange's revenue, which turns to reflect the revenue of the whole advertising ecosystem. Recall that online advertising is said to be supporting the current Internet free business model. Thus, at least for now, this kind of solutions should not significantly tamper with the current ad distribution model since it could negatively affect the economy of online advertising platforms.

4.4.2 Results

We set up a scenario with twenty DSPs: seven bidding high, seven bidding low, and six bidding between high and low (an average value). Then we simulate an ad exchange instance holding a thousand auctions. Our distribution control strategy is

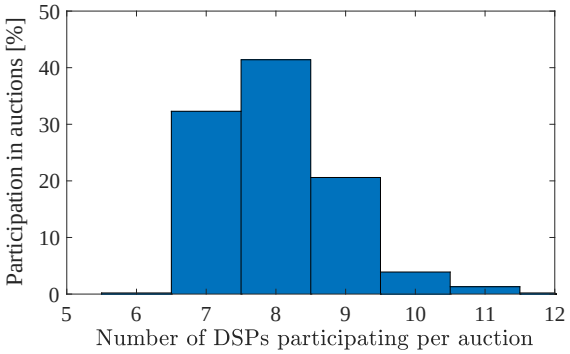


Figure 4.9: Number of bid requests (invitations to DSPs) sent per auction for our experiment, with $\alpha = 8$.

enforced with $p_{\min} = 0.05$ and with $\alpha = 8$. That is to say, to prevent abuses and preserve privacy, bid requests are distributed among eight DSPs in average (those with better behavior) and not among the twenty available. Furthermore, a minimum of 5% of bid requests are distributed among these eight DSPs in order to guarantee all them will participate at some point.

In Fig. 4.9 we represent the number of DSPs, out of the 20 available, that participate in each auction of our experiments. As expected, this histogram confirms that the number of DSPs participating per auction is 8 in average (the value of α).

Then, we also use Fig. 4.10 and Fig. 4.11 to characterize the participating DSPs in terms of the rate of won auctions (ω) and the average money spent (μ), respectively. These parameters are measured at every auction, with respect to all the previous auctions. We depict the values to describe the behavior of three DSPs, one from each category. Evidently, these figures show how DSPs with a more desired behavior (bidding higher or spending more) present better indicators ω and μ.

Additionally, we assess the effects of our mechanism on the revenue of the ad exchange. For this, we perform a set of experiments using different values for the parameter α, from 1 to 20 (i.e., we simulated a round of 1000 auctions for each value of α). As α represents the average number of DSPs to which bid requests are sent from the ad exchange, the results from our experiments reveal the impact of the value of this parameter on the total revenue obtained. This impact is illustrated in Fig. 4.12, for the two different strategies for generating bid requests (uniform and

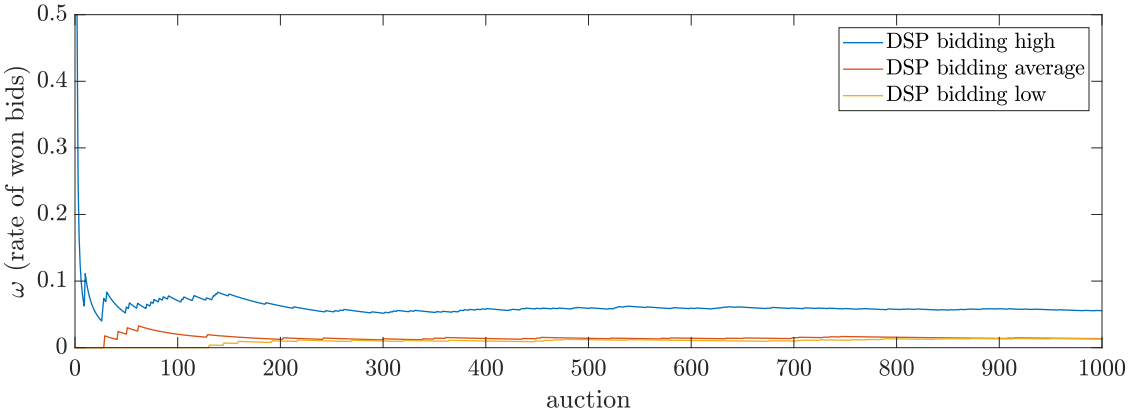


Figure 4.10: Rate of won bids for different DSPs behaviors in our experiment. We use $\alpha = 8$ and $\lambda = 0.05$

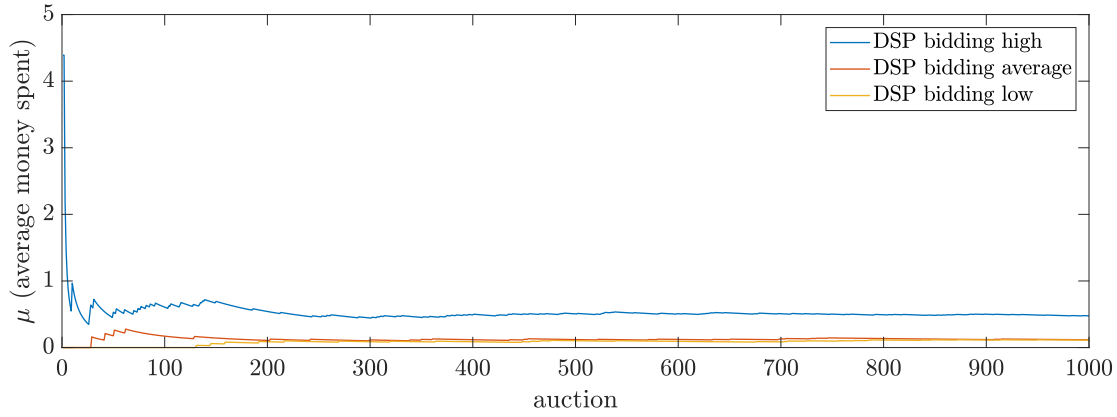


Figure 4.11: Average money spent according to different DSPs behaviors in our experiment. We use $\alpha = 8$ and $\lambda = 0.05$.

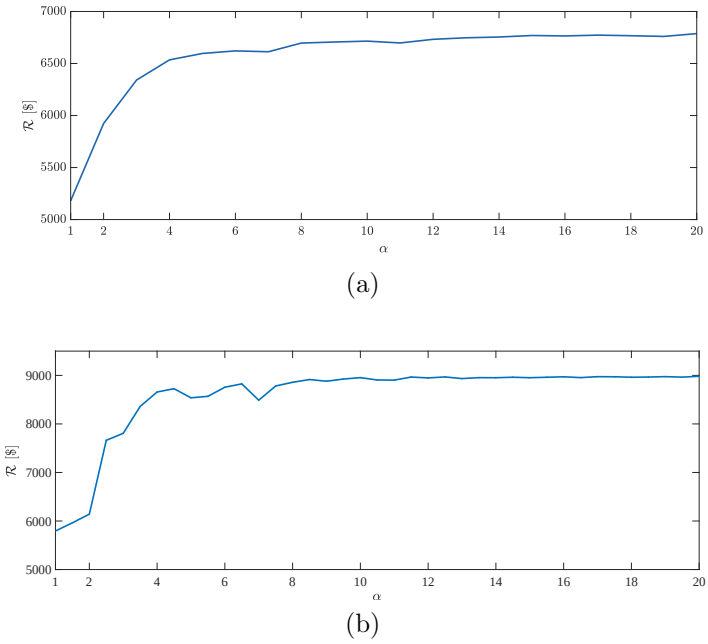


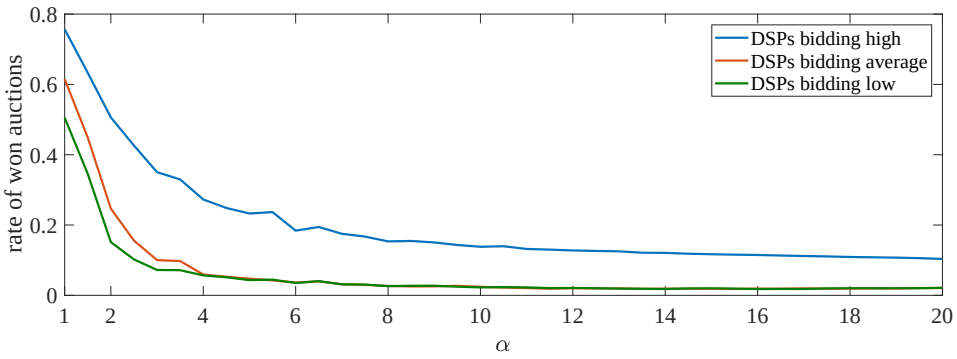
Figure 4.12: Revenue obtained by the system for different values of α. For these experiments we use $\lambda = 0.05$ and a set of 20 DSPs so we make α vary from 1 to 20. Experiments are made following two different random distributions when generating the bid requests: (a) uniform and (b) Gaussian.

Gaussian distributions). First, note how the revenue increases with the value of α, consistently with the tradeoff commented in Sec. 4.3 and depicted in the conceptual plot in Fig. 4.8. In addition, when $\alpha = 20$, the maximum revenue is reached because, in practice, no control mechanism is applied when all the available DSPs are activated to receive bid requests. Remarkably enough, the revenue when $\alpha = 8$ and onwards is pretty close to the revenue when $\alpha = 20$. Actually, in those cases, revenue is less than 1% lower than the maximum obtained when our strategy is not applied. The importance of this result lies in that the bid request distribution control enables certain privacy guarantees that can be enforced while having a very small impact on the ad exchange's economic benefit. The results observed in these experiments, however, are certainly tied to the specific behaviors assumed for the DSPs. As a matter of fact, our theoretical analysis of the trade-off between revenue and data distribution control found that $\mathcal{R}(\alpha)$ depends on the sequence of slopes ω_i μ_i. The higher the slopes of the first DSPs (sorted according to (4.2)), the fewer the average number of DSPs needed to obtain revenues close to $\mathcal{R}_{\max}$. On the extreme, the case

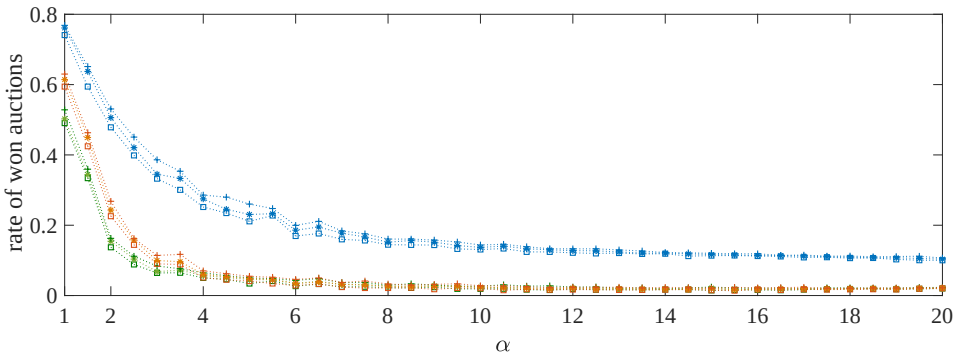
$\omega_i\mu_i = \omega_j\mu_j$ for all $i, j = 1, \ldots, d$ yields a straight line, which represents the worst scenario in terms of ad revenue.

Finally, we are interested in seeing how our parameter α is capable of regulating the behavior of DSPs. For this, we conduct an experiment with a setup of 3 DSPs, each behaving differently (bidding high, low and average). We simulate a thousand auctions and apply our privacy mechanisms for different values of α, from 1 to 3. We measure the rate r of won auctions with respect to the requests (invitations) received by each DSP from the ad exchange. This rate could be interpreted as a measure of the goodness of the behavior of DSPs as stated in Sec. 4.3. A DSP bidding higher shall win more bids and spend more money. Accordingly, the higher the rate of won bids, the more desirable is the behavior of a DSP. Conversely, r could also be seen as a measure of the abuse committed by a DSP against the privacy of a user, since a low rate of won auctions (low r) would entail a DSP receiving user data information without paying for it.

The results of this experiments are illustrated in Fig. 4.13 and Fig. 4.14, where we depict the evolution of the rate of won auctions of different types of DSPs. Respectively, we plot the results obtained from using two different strategies to generate bid requests (uniform and Gaussian) for each type of DSP. First note that, in this context, $\alpha = 20$ represents the case where the ad exchange sends bid requests to all available DSPs, so there is no control strategy applied. In this case, we see that DSPs bidding low have low rates of won auctions, which would imply that they are taking advantage of the advertising system. However, if we analyze the value of this rate as α decreases, we observe that the rate r increases for each type of DSP, which suggests a successful adjustment of the behavior of DSPs. In general, thus, it makes sense to maximize the benefits of the ad exchange subject to a restriction by distribution control (privacy) since the rates of won bids shall improve for small α.

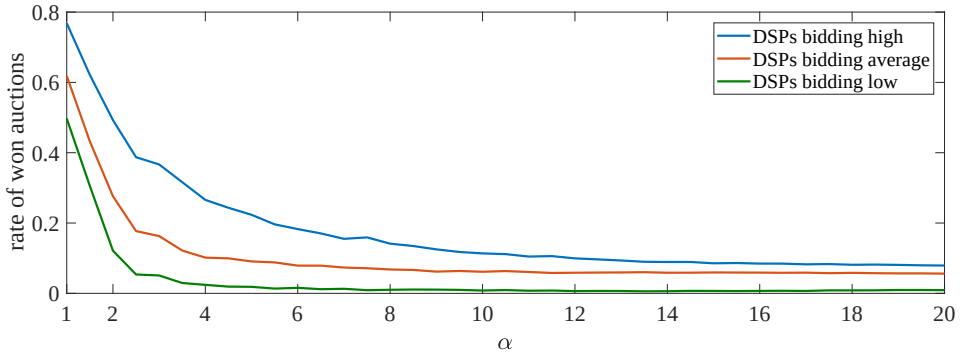


(a)

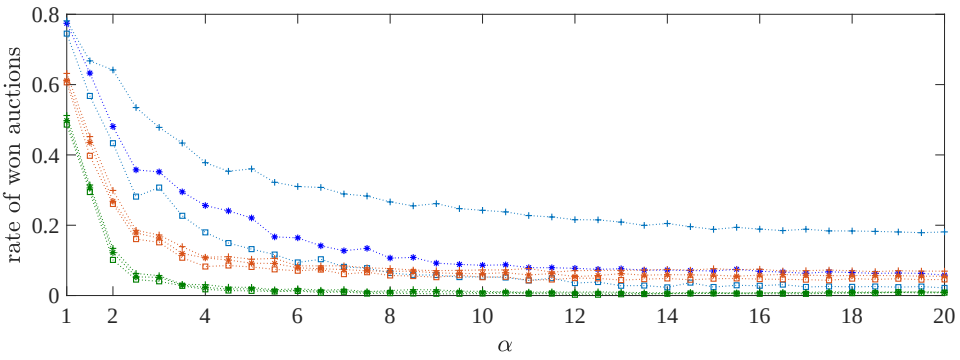


(b)

Figure 4.13: Evolution of the rate of won auctions for different values of α (from 1 to 20 in steps of 0.5) and $\lambda = 0.05$. We consider 20 DSPs with different behaviors (bidding high, average and low). For each value of α, we repeat the experiment 20 times. The results are depicted averaged in (a) for each type of DSP. In (b), we illustrate results of percentiles 3 (using '+'), 50 (using '*') and 97 (using '□'). For these experiments, bid requests for the different types of DSPs are generated using a uniform distribution.



(a)



(b)

Figure 4.14: Evolution of the rate of won auctions for different values of α (from 1 to 20 in steps of 0.5) and $\lambda = 0.05$. We consider 20 DSPs with different behaviors (bidding high, average and low). For each value of α, we repeat the experiment 20 times. The results are depicted averaged in (a) for each type of DSP. In (b), we illustrate results of percentiles 3 (using '+'), 50 (using '$*$') and 97 (using '$\square$'). For these experiments, bid requests for the different types of DSPs are generated using a Gaussian distribution.

4.5 Discussion

The big picture of privacy in the online advertising ecosystem

The "hyperconnection" of people to the Internet is making them widely traceable by the providers of third-party applications that enable the collection of personal data. Among such providers we find online advertising platforms, which might be building a mass surveillance structure due to several reasons. First, the presence of advertising online is so *massive* that both the tracking of users and the collection of their data are real-time and ongoing processes. Namely, personal data is continuosly leaked to the advertising infrastructure as users browse the Web.

Besides, in this same direction, we have verified that the *reach* of advertising entities is pretty wide. It is a fact that advertisements "follow us" wherever we surf the Internet. As a consequence, user information is regularly processed in bulk and indiscriminately, always in the name of greater personalization. Furthermore, the user information ceded to third parties during ad auctions is extremely *granular*. This facilitates identification of users, e.g., by using identifying attributes (such as IP addresses) or combining them to build a fingerprint. Due to granularity, not only identification is feasible, but also other privacy attacks derived from the type of information released. For instance, variations in location data along with IP address changes could unveil user movement patterns. Also, information about sites visited may reveal the interests and behavior of users. Finally, these practices of ubiquitous tracking and aggregation of granular user information is largely concentrated in entities over which little control is enforced. Sadly, this concentration of the power of surveillance is not only affecting the privacy of many users but it is turning advertising entities into dangerous means of massive manipulation, as exemplified in [146].

This scenario, which is less and less encouraging for privacy, is made worse by the lack of transparency in the sharing of user information among the participating entities. Hence, users are unaware of the complex dynamics behind the advertising ecosystem and the particular privacy risks they are facing online. And so forth, partly motivated by some creepy perceptions regarding online behavioral advertising [69], people are increasingly using ad blockers. Whilst emerged to undermine abusive

tracking from advertisers, ad blockers bear an interesting concept in giving users a more active role in the advertising ecosystem. This role might consist in providing users with more transparency and radical control over ads. However, very little can be really done if abusive behaviors that exacerbate privacy risks are ignored within the core of advertising platforms.

The privacy risk derived from user data sharing

One of the abusive behaviors that threat privacy in the online advertising ecosystem involves the malicious collection of bid requests by DSPs and related third parties who, violating the terms of service defined by ad exchanges, may be participating in auctions without any interest in winning. This is possible due to the oversending (broadcasting) of bid requests including personal data to DSPs, which is motivated by the need for an ad exchange to maximize profit.

Regulating bid request distribution as a mechanism to preserve privacy

Our contribution to address this issue consists in enabling some control over a crucial part of the advertising ecosystem: the bid request distribution model to DSPs and similar intermediaries. In this line, our experimental results show that reducing the number of DSPs recipients in online advertising through regulating such distribution may virtually cause no losses in revenue for the ad exchange. Particularly, the higher the value of the parameter r_i (product of the rate of won bids and average money spent) for the DSPs with better behavior, the less DSPs need to be contacted to reach the maximum revenue; thus, personal information would be shared among less third parties. Consequently, based on this control strategy, DSPs are encouraged to correct their behavior since, otherwise, their chances of participating and winning in ad auctions fall significantly. Despite the small losses in revenue, we state that win/win outcomes are reached for the interests of ad exchanges (revenue) and end users (privacy) since *actively regulating the behavior of third parties regarding*

user privacy could significantly discourage harmful attitudes of users towards online advertising (e.g., massively using ad blockers).

Unlike other approaches to preserve privacy in online advertising, which contemplate significant modifications in the ecosystem, a great added value is provided by ours since it entails a *minimum change in the bid request distribution strategy*, while leaving the main online advertising infrastructure untouched, albeit personal information is still ceded to third parties. This should be a great incentive for ad exchanges to adopt this kind of mechanisms in order to regulate the behavior of associated agencies and to take additional steps to protect the privacy of end users.

Interestingly, our approach could be extended to complement transparency and control enabled in the user side through an interface, e.g., the one offered by an ad blocker plugin such as Adblock Plus. First, an ad exchange implementing our bid request distribution model might provide users (through this user interface) with the value of α, i.e., the number of entities with which their personal data has been shared (transparency). Furthermore, as a privacy mechanism, the browser plugin could enable users to configure a maximum number of entities with which to share their data (informed decision). Accordingly, if the user data results to be shared with more entities in average, the plugin would block the corresponding ads (control).

In the same line of reducing the potential adversaries to protect privacy, an improvement of our approach could be *targeted auctioning*. This would consist in partitioning our optimization problem to be solved on a per-market basis, i.e., auctioning a user impression only among the DSPs subscribed to a given targeting market. Said otherwise, the specific targets of a DSP at a moment in time could serve as another reputation parameter when the ad exchange auctions a user impression.

Appealing to a change in the bid request distribution model, in the core of the advertising ecosystem, entails a big step towards enforcing privacy in this context; more if the impact of such strategy can be minimized. As depicted in the previous paragraph, although bringing some controlled loss in revenue, our proposal may suggest a paradigm shift with a multiplicative effect for the benefit of user privacy. Not only is activated a technology for ad exchanges to support privacy regulation, but part of the control can be given to users. And further, this approach could serve

to alleviate the harmful tensions between advertising systems and users provoked by serious concerns regarding privacy.

Privacy protection with our system

The extent to which our mechanism could protect privacy may also be subject to discussion. Whereas the level of privacy provided by some mechanisms could be quantitatively measured under certain assumptions, whether the given protection is sufficient or not is pretty relative. This is because, in general, the level of privacy provided by any protection mechanism depends on the context, and in our case, it is defined by the requirements set in Sec. 4.3.1, the adversary model from Sec. 4.1.1, and the strategy proposed in Sec. 4.3.2. In this specific framework, our solution could provide great privacy by enabling an ad exchange to strongly support privacy without significantly affecting the revenue of the system. However, the ultimate level of privacy provided would depend on the particular strategy adopted by the ad exchange (either, e.g., aggressive, capping a lot the participation of DSPs; or moderate, not restricting it significantly).

Beyond this limited scenario, other players might still disclose user information, e.g., first parties (publishers), ISPs, data brokers, etc. However, the scope of action of ad exchanges is by far greater. Since DSPs may illegitimately benefit from such capabilities, extremely reducing the amount of DSPs participating in auctions, e.g., to a dozen, would improve privacy to a similar extent.

In any case, ours is a first approach to dealing with privacy issues in this particular context where user data may be inappropriately shared to dishonest DSPs. Interestingly enough, future work might concentrate on giving users further control capabilities focused on modulating the privacy parameter introduced in this strategy. Thus, if provided with some background information, users themselves would be able to choose the privacy level they feel comfortable with.

Incentives to adopt regulated data distribution

Unlike most proposals to protect privacy in the online advertising ecosystem, ours aims to encourage advertising players (mainly ad exchanges) to adopt it for the benefit of users and ad platforms. In this subsection, we describe the main incentives that these entities would have to implement the mechanism presented in this chapter.

- Privacy regulations require more and more control from data controllers and processors over user data collection, use, and distribution [65]. Furthermore, the heavy fines for neglecting user privacy are pushing these entities to adopt protection mechanisms, especially when they manage user data at a massive scale [147]. Since pur approach aims at the more private distribution of user data at a relatively low cost, we think that ad exchanges would be strongly motivated to adopt it.

- Not only regulation is urging the advertising ecosystem to endorse privacy protection initiatives, the current ad blocker arms race [148] is empowering users to protect themselves through radical mechanisms that might be affecting the economic health of the advertising players. In such a conflictive environment, it seems reasonable for these actors, especially those in the core network, to start to give in to the users' legitimate expectations of privacy. Otherwise, the war would grow fiercer, seriously affecting not only the online advertising business model but that of the entire Internet.

- The implementation of the distribution mechanism we propose would not involve significant changes to the online advertising ecosystem. The ad exchange would only have to incorporate a module to distribute bid requests among the "best behaved" bidders for a given context (privacy parameter, targeting market, etc.). The rest of entities would remain unchanged.

- Even though our approach does not tamper with the current online advertising model, it might generate a regulation effect over DSPs. That is, in order to avoid penalization, DSPs would not adhere to dishonest practices when participating

in auctions. Interestingly, this value-added service could further improve the system's revenues.

- As it has already happened with other privacy-enhancing initiatives (e.g., Facebook's ad preferences), ours opens the door to the implementation of further transparency and control mechanisms for users. Our mechanism would encourage those ad exchanges committed to respect user privacy to create interfaces for users to examine or even modulate the privacy parameter we are introducing here.

- Beyond its technical implications, ad exchanges would be highly encouraged to implement our proposal in order to compensate users for the opacity behind which the exploitation of their data has been hiding. The scandals surrounding the abuse of user information undermine daily the reputation of the advertising ecosystem and hence the trust of end users, the ultimate owners of data (the main input of the online advertising ecosystem). Upon realizing that specific controls are being implemented to protect their privacy, their concerns could be alleviated, since their main concern is not the sharing of information itself, but the inappropriate sharing of their information [13]. Consequently, users themselves could even decide not to block the tracking of privacy-compliant ad exchanges, which is a further incentive for the latter to adopt our mechanism.

4.6 Conclusions

Undoubtedly, the main privacy concerns regarding online advertising come from the great capability of third parties to aggregate user data. Due to the inherent opacity of this ecosystem, the most known approaches to face such concerns build on radical ad blocking solutions. By entirely blocking ads and partly stopping the leakage of data from the user side, these radical approaches are threatening the current economic model of the Web. On the other hand, with the aim of balancing the trade-off between revenue and the number of invited DSPs (looking for more user privacy), we propose to modify part of the ad delivery model. Our technique arises as a strategy

of bid request suppression where interactions carrying user data can be reduced, by design, to offer more privacy, while slightly affecting the revenue of the system. More specifically, we come up with a controlled distribution of bid requests among DSPs in order to reduce the amount of user data shared with said third parties. Nevertheless, our approach comes at the expense of revenue loss incurred by lowering the number of participants within ad auctions. Since this technique would be applied directly in the core of ad platforms, more overwhelming and less harmful results could be obtained.

Part of our effort lies on an analysis of the privacy risks involved in the massive aggregation of data performed by some online advertising entities. In this line, we strove to characterize the personal information leaked in bidding interactions and some of the derived critical jeopardies. We concentrate on bid request messages that are used to invite DSPs to participate in ad auctions and that carry very granular information about the user online behavior. Thus, using a publicly available data set belonging to a famous chinese DSP, we unveil the potential capability of advertising intermediaries to do massive surveillance even at a very low cost. Accordingly, we also highlight the power given to advertisers to microtarget users with a very fine precision.

Our main outcome is a mathematical approach to tackle the aforementioned problem of distributing bid requests to less DSPs, while minimally affecting the revenue of the system. We formulate and solve an optimization problem that seeks to maximize the revenue while bounding the participation of DSPs. Thus privacy is enforced through balancing this revenue-invitation rate trade-off.

As a result of our theoretical analysis, we present a close-form solution for the bid request distribution strategy and a revenue-invitation rate function characterizing the optimal trade-off curve. From this analysis, we find an interesting opportunity to cap the number of DSPs that receive bid requests while maintaining a reasonable revenue. From simulations perfomed over an auctioning scenario, we confirm that the revenue of the system indeed increases with the number of DSPs participating in each auction. However, we find that even when drastically reducing this number (thus, increasing privacy of users) an important portion of revenue may still be preserved. Also, it turns out useful to maximize revenue subject to a restriction that supports

privacy when handing out bid requests, because it leads DSPs to behave better (e.g., increasing their rate of won bids), driven by a penalization on abusing the system (e.g., when bidding too low).

Chapter 5

Measuring Third-Party and Advertising Tracking in Iberoamerica

5.1 Introduction

Personalized online advertising is responsible for much of the online tracking performed over users. Online advertising platforms are supported by sophisticated personalization systems that tailor ad content according to the preferences of users; these preferences are learned from the information collected by tracking. In this line, the more information is collected, the better the performance of personalization systems, and the higher the profits of the advertising platforms. Since online advertising has become a millionaire business [149] that apparently supports the very existence of the Internet [150], there is a great motivation from multiple instances to collect more and more data, which implies massifying and improving online tracking.

Online tracking refers to the activity of closely following a user wherever she "goes" while browsing the Web. This is possible because users leave innumerable footprints online, without even noticing it, when requesting for content to websites. IP address, operating system, browser type, plugins installed, patches applied, and browsing history are some examples of (context) information leaked in a single HTTP

request. If aggregated and processed, said information could serve to build user profiles revealing location, shopping habits, entertainment preferences and even the gender of users.

With the involvement of so much user information, online tracking raises serious privacy concerns. The information collected may be so varied and detailed (e.g., location, interests, voting preferences) and the technology used so specialized that tracking may enable external entities to characterize a significant part of a user's life. Furthermore, state data is currently being collected, due to a real-time mechanism that binds online tracking with every single user web request, enabling third parties (not only Internet providers) to literally monitor each of the user "movements" on the Web.

In this attacker scenario, the first potential tracker is thus the website (publisher) that the user visits, as described in Sec. 3.2.1. Thus, if tracking is performed from the publisher, it is called *first-party* tracking. In general, the audiences of first parties are pretty segmented, so the user tracking they might perform is usually innocuous. Some exceptions are the 'walled gardens' built by the Internet giants (e.g., Facebook), which concentrate services for millions of users within a single ecosystem.

Furthermore, a single user web request commonly triggers connections from the user browser to several *third parties* that receive part of the aforementioned contextual information. This information is used by third-parties to support real-time services such as personalized advertising or other services for websites, e.g., media hosting (by content distribution networks), load balancing, or social networking. Figure 5.1 illustrates the interactions triggered by a user browser request, which enable first and third-party tracking.

Undoubtedly, better online services are provided thanks to personalization and outsourcing; however, third party tracking supports the massive aggregation of user information (collected along multiple sites along the Web) in the hands of anyone aiming at paying for it [23]. The online activity of users is then monitored and processed by several entities that users had never heard of. Figure 5.2 illustrates the large number of connections to third parties (information flows) derived when a user visits only three websites.

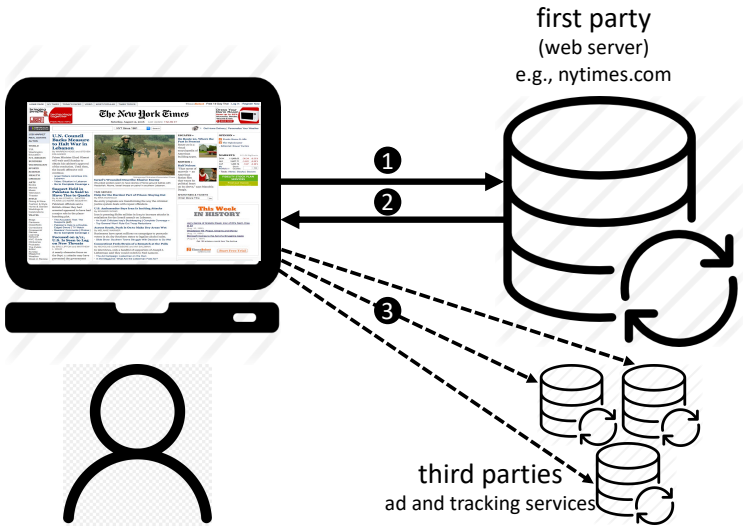


Figure 5.1: Requests to third parties (3) triggered by a single HTTP user request (1). When a user browses a webiste, a redirection command is commonly sent in the HTTP response (2) to spawn further connections to third parties.

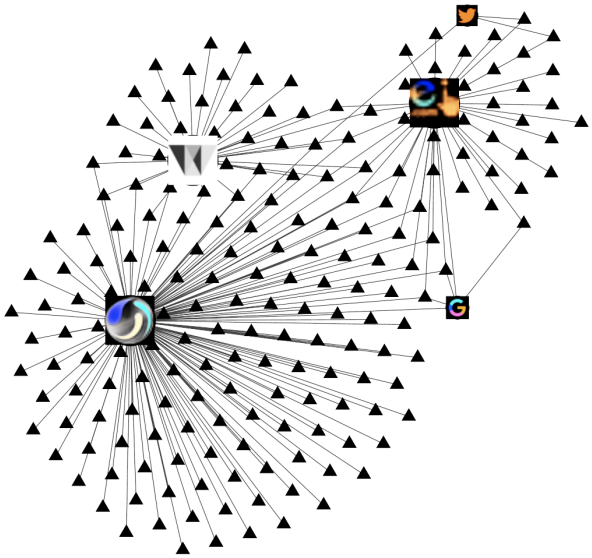


Figure 5.2: Illustration of the multiple connections to third parties (more than 50) generated in the background after visiting only 3 sites. The points where connections originate represent the websites while the little triangles represent the third parties contacted. This figure was obtained through the browser extension Disconnect [1].

Unlike physical user tracking, e.g., in a street, online tracking is much less evident for users because it is performed on the background of web browsing. Unfortunately, there is very little evidence available for users to realize the latent pervasiveness of online tracking.

Due to its prevalence on the Web, measuring online tracking could be a way to characterize the privacy risks of Internet users. The severity of said risks may be illustrated through different indicators such as the level of exposition of user interactions to third parties, the concentration of user information on a few advertising companies, the dynamic behavior of tracking for websites belonging to certain categories, or the suspicious requests to third parties triggered when accessing government web sites.

While related work [81] has performed more general approaches by studying online tracking through the most popular sites of the Web, our analysis in this chapter was mainly devoted to assess the online tracking and advertising triggered by local websites, in particular those from Iberoamerican countries. Moreover, our effort was oriented to find out how the location of users and the type of publishers impact on the online tracking and advertising interactions and, thus, how on the potential privacy risks. Since most related research has focused on top-world, American, or European sites, probably, other aspects not addressed in said work will arise from the work presented in this chapter. We are interested in a very heterogeneous region such as Latin America (LATAM) and its contrast with the European Union (EU), e.g., in terms of the adoption of modern personalization systems (including online advertising) or the maturity of user perceptions regarding privacy.

Our assessment was also based on taking advantage of the `ads.txt` standard, increasingly being adopted by websites for some years now. `ads.txt` is a project promoted by the Internet Advertising Bureau (IAB) to increase transparency in the programmatic advertising ecosystem and prevent fraud. It encourages publishers to publicly inform the companies they have authorized to sell their advertising inventory (ad spaces). Such publication is done through a text file (so much like the `robots.txt` standard) called `ads.txt` in the root context of the website. Interestingly, revealing

such information could also serve as a transparency mechanism for users so we collected and processed the content of this file to confirm the results obtained when crawling third-party tracking.

The work presented in this chapter is related to that published in [18].

Chapter outline

The rest of this chapter is organized as follows. Section 5.2 describes the methodology followed in this work, including data collection, processing, and experiments. Section 5.3 presents the results of measuring online tracking and advertising in Iberoamerica. Finally, conclusions are drawn in Sec. 5.4.

5.2 Experimental Methodology

We study the impact of online tracking and advertising in Iberoamerica by measuring the third-party traffic triggered when browsing websites in this region. In the next subsections we explain how the data related to this traffic was generated, collected and processed. Also we describe which countries and websites were involved in this analysis. More importantly, we briefly describe the tests performed to study the privacy risks in this particular context.

5.2.1 Selection and Categorization of Websites

Selection of Websites

Since the context of our study builds on Iberoamerica, we first selected the countries whose websites would be considered for our tests. Thus, we selected the countries in this region that allowed us a VPN connection so that web traffic could be generated from such locations. This included Spain and Portugal from EU and the following LATAM countries: Argentina, Brazil, Chile, Colombia, Costa Rica, Ecuador, Mexico, Peru, Uruguay, and Venezuela. For some tests involving web traffic directed *to* LATAM, we also included other countries.

Next we dedicated to obtain the most popular websites within each of the aforementioned countries. For this, we manually selected only the local websites from the top 500 ranking published for each country by the Alexa Top Sites service offered by Amazon [151]. For other experiments, we also collected the top 500 global websites also using this service as source. In a nutshell, our experimental scenario consisted of 12 countries, 2,076 Iberoamerican websites and 500 top-world websites.

Categorization of Websites

With the aim of understanding the influence of the content published in websites on the phenomena studied in this work, we also labeled each website tested according such content. We manually categorized each of the websites, with the support of the Site Safety Center tool of Trend Micro [152].

5.2.2 Data collection

The data we studied mainly included all the third-party requests spawned after visiting a website. Assessing the magnitude of such traffic, the destination third-parties participating, and even the tracking information they may set in the user side (cookies), may help to unveil the inherent privacy risks of users browsing these sites.

Our experiments involved simulating websites visits to trigger third-party traffic. For the thousands of websites in our scenario, we performed this (including data collection) automatically through OpenWPM, a very versatile tool devoted to web measurement [81]. OpenWPM offers a programmable interface to orchestrate the main functions of a web browser, thus allowing automated web crawling and collection of tracking-related information (redirects, cookies and third-party calls) that is stored in a SQLite database. This tool was also used to obtain, if existing, a particular file from websites, to study the adoption of the `ads.txt` advertising standard.

5.2.3 Experiments

We performed several tests simulating web traffic from different countries in Iberoamerica, given our particular interest in studying potential privacy risks derived from the

location of users when browsing sites in this region. As suggested previously, we used a VPN service to connect to each of said countries and send web traffic to local websites.

First, we generated *local traffic*, i.e., visits from each country to websites in the same country. Furthermore, we simulated visits from (a country in) EU to websites in LATAM countries. This way we tried to detect different effects when the same publishers are visited from diverse locations. But a particular website, especially if widely popular, seems an important parameter of third-party traffic too. Thus, for the sake of comparison, we also experimented simulating web traffic from Iberoamerican countries to top-world sites.

Through OpenWPM we collected information on the third-party requests triggered by simulated user browsing, and on added third-party cookies set on the user side. From said data, we also measured advertising related requests, which, associated with a personalized service, may imply higher privacy risks. To measure the dynamics of advertising interactions, we classified third-party requests as ad related or not, by resorting to available libraries that, based on ad blocking lists [54], facilitate the detection of such type of traffic.

We aimed at finding privacy risks by measuring the intensity of third-party traffic (including ad related interactions) and counting the number of entities behind such traffic, and the cookies set in the browser. Third-party traffic, measured as the mean number of third party requests triggered from websites, is a first approach to unveil potential privacy issues. The greater the indirect and non-consensual traffic from users the more user information would be released through such flows. In the same line, we obtained the number of third-parties receiving said traffic as a proxy to the number of potential third-party trackers. We also counted the mean number of cookies set by third-parties, and particularly those cookies more likely to be related to user pseudo-identifiers. Since online tracking tightly depends on this information, measuring this parameter contributed to our objective of finding privacy risks.

To ascertain the impact of the type of content served by publishers on the intensity of third-party tracking, we tabulated the information collected according to the categories of websites. Since the consumers of certain (categories of) content might

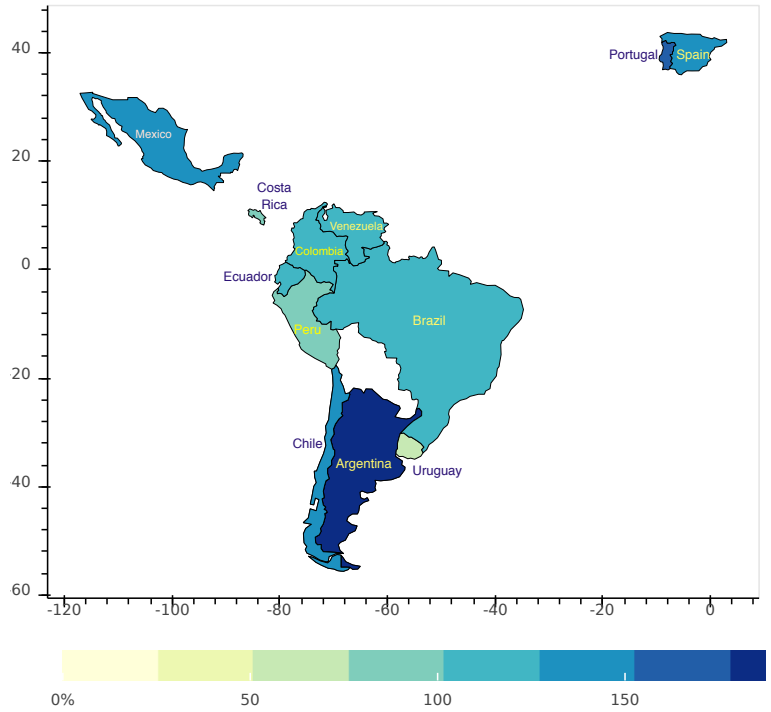


Figure 5.3: Heat map illustrating the mean number of third-party requests triggered from local traffic in Iberoamerican countries.

be more relevant in economic terms, the sites associated with such content might be more exposed to third-party tracking and privacy risks.

Finally, trying to confirm some of the findings when studying third-party tracking, we collected the records in `ads.txt` files from Iberoamerican publishers where this file was present. From the data obtained, we studied the adoption of this standard in Iberoamerica and the monopolistic influence of some companies on the advertising ecosystem.

5.3 Experimental Results

5.3.1 Third-party Traffic

In Fig. 5.3, we illustrate the impact of third-party traffic triggered by browsing Iberoamerican web sites locally, i.e., from each country. Such map representation enables to illustrate the geography, size, and potential population of involved countries.

This perspective reveals the marked heterogeneity not only among Latin American countries, but particularly among them and EU countries (here represented by Spain and Portugal).

A lot of traffic towards third-parties is observed in this context. Besides the great number of third-party requests triggered in general by a single visit to publishers, we evidence that such traffic is not homogeneous along countries. It stands out that this traffic within some large and populated countries, such as Brazil and Peru, has a lower impact than that in Portugal and Spain, although the latter are much smaller and less populated. On the other hand, Argentina, Mexico, and Chile do show an important number of third-party requests spawned by local traffic.

Beyond the disparity in Latin American countries, European publishers and users seem to be more attractive targets for third-party tracking. This might imply a higher risk for such users since more entities and more personal information would be involved. However, many of these third-party requests could have the same destination, so it is convenient to identify the recipient entities (third-party trackers) by filtering the domain names from their destination URLs. As noted above, insofar these entities receive so much indirect user traffic, they might become attackers, not only against user privacy but also against user security.

As commented in Sec. 5.1, the number of third-party domains behind online tracking may better reflect a potential privacy attack scenario since it could serve as a proxy of the number of third-party entities collecting information (third-party trackers). As depicted in Fig. 5.4, hundreds of entities were found receiving third-party requests, indirectly, from users, when locally visiting Iberoamerican sites. European countries had a very similar number of potential third-party trackers while in Latin America the situation is less uniform. In Brazil and Mexico, by far the most populated countries, we found the greatest number of third-party entities. However, we think the difference with other countries is certainly minor, considering the variation in population.

One might think that websites in a country of hundreds of millions of citizens, such as Brazil, would spawn much more third-party traffic than those in a country with a couple of tens of millions (e.g., Chile, Spain, and particularly Portugal), but

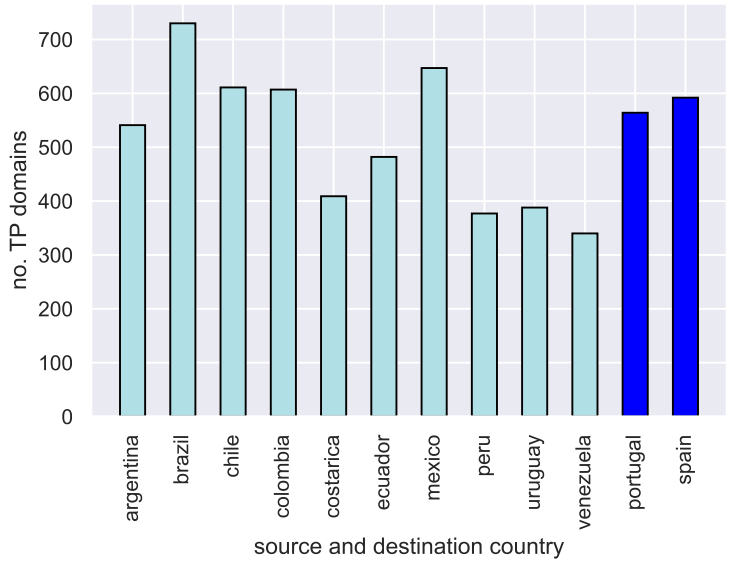


Figure 5.4: Number of third-party domains contacted as a result of local traffic within Iberoamerican countries.

the difference is not as marked as the population. There might be some reasons. First, it could reflect a preference of potential trackers to certain type of population. However, third-party requests are also generated by technology services, e.g., content distribution, commonly used by publishers. Thus, more entities behind third-party traffic could also indicate more tech supporting websites. In any case, there would be more privacy risk for a user if he is targeted in a small population than in a big one.

By weighting the number of third-party domains by the population of countries (in millions of inhabitants) as shown in Fig. 5.5, we tried to capture the latter effect. We can see in this figure that, despite being very sparsely populated, Costa Rica, Uruguay, and Portugal spawn a lot of third-party traffic. Despite its size, these are relative rich countries, in particular when compared with the average in Latin America, which we think could explain this behavior to some extent. Paradoxically, richer countries are more prone to implementing strong privacy legislation, which in this context does not seem to discourage third-party traffic. Note that this behavior is measured when crawling local sites of each country *from* the same country.

Trying to identify third-party tracking, and specifically personal data leaking, another interesting indicator might be the mean number of third-party cookies set

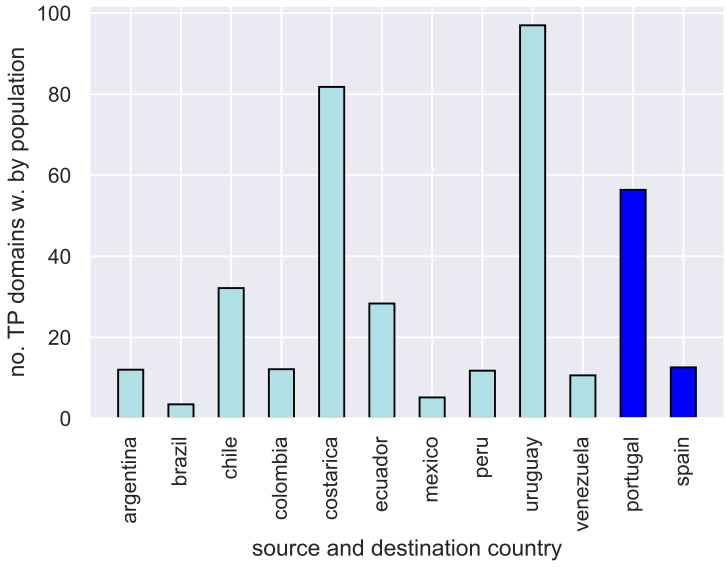


Figure 5.5: Number of third-party domains contacted as a result of local traffic, weighted by population, within Iberoamerican countries.

in the user's browser. Recall from Sec. 3.2.2 that third-party cookies may be used to transport user identifiers that online trackers employ to recognize a user when she visits a publisher. This way, a tracker is able to "follow" users and associate information to their profiles. We cataloged these cookies as tracking cookies or identifying cookies (ID cookies) if their lengths were greater than 6 as done in previous works.

When processing the data obtained from local traffic within Iberoamerican countries, we found that a user browsing local sites from Portugal would receive 14 ID cookies on average, as depicted in Fig. 5.6. Spain, Argentina, and Colombia follow with 8 ID cookies on average. This figure shows that this number varies along the different countries but suggests a great interest of Portuguese local sites in tracking local users.

Since we categorized each of the publishers visited along our experiments, we could represent in Fig. 5.7 the potential influence of the content delivered by publishers on the tracking performed over users. As shown in other works, `news/media`, `entertainment`, and `shopping/travel` are the categories concentrating more third-party tracking, in this case when local traffic is studied in Iberoamerican countries. Some particularities should be noted about some countries: users from Costa Rica

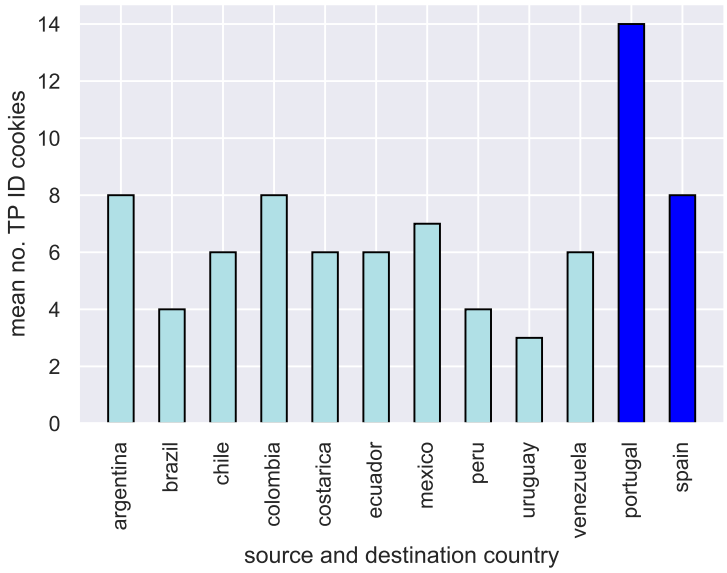


Figure 5.6: Mean number of identifying third-party cookies found as result of local traffic within Iberoamerican countries.

would be receiving a lot of third-party tracking triggered by education and government sites; and, from Brazil, would occur the same with `weather`.

As a consequence, some of these categories may entail more privacy risks than others. For instance, `education` could involve audiences including children or teenagers who are clearly more vulnerable to abuse, more if the sites belonging to these category require any kind of login. Also, probably, government sites should not be tracking their citizens, even worse to "profit" from their interactions if coupled, e.g. with advertising platforms.

When comparing traffic to LATAM originated locally vs. from EU, traffic from EU spawns, in general, more third-party requests per website as shown in Fig. 5.8. Namely, when traffic originates in EU, more third-party interactions are observed in most of LATAM countries. Chile is the exception.

Although the intensity of third-party traffic is higher for web browsing from EU than internally in LATAM, the total number of domains or entities behind third-party traffic is very similar, no matter where the web traffic originates, as depicted in Fig. 5.9. The same entities would be involved in third-party traffic but changing the tracking strategy depending on the location of the user.

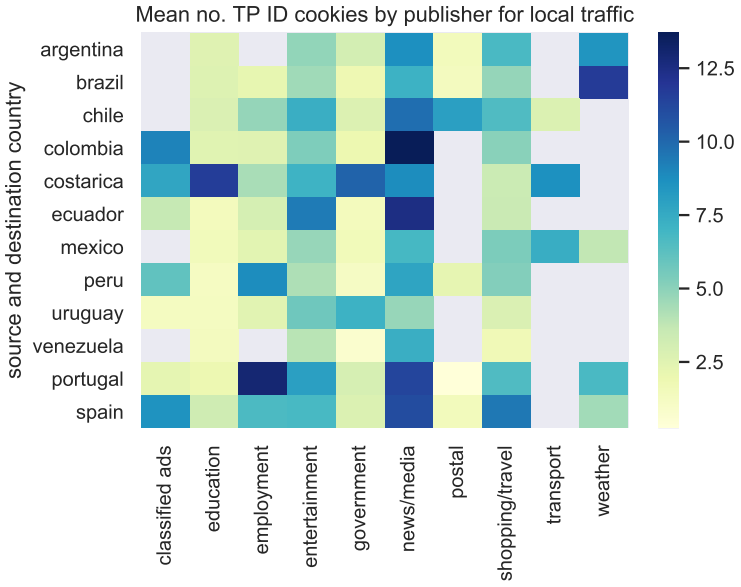


Figure 5.7: Mean number of ID third-party cookies set from local traffic within Iberoamerican countries, organized by category of publisher.

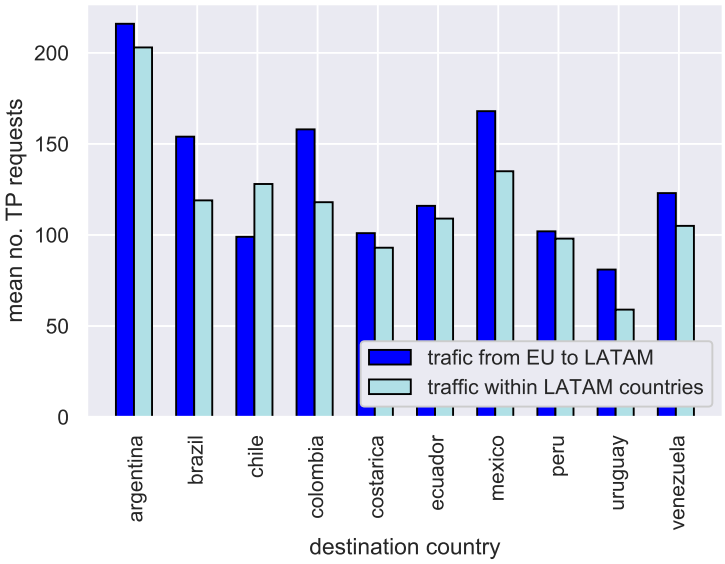


Figure 5.8: Mean number of third-party requests triggered by web traffic from EU to LATAM.

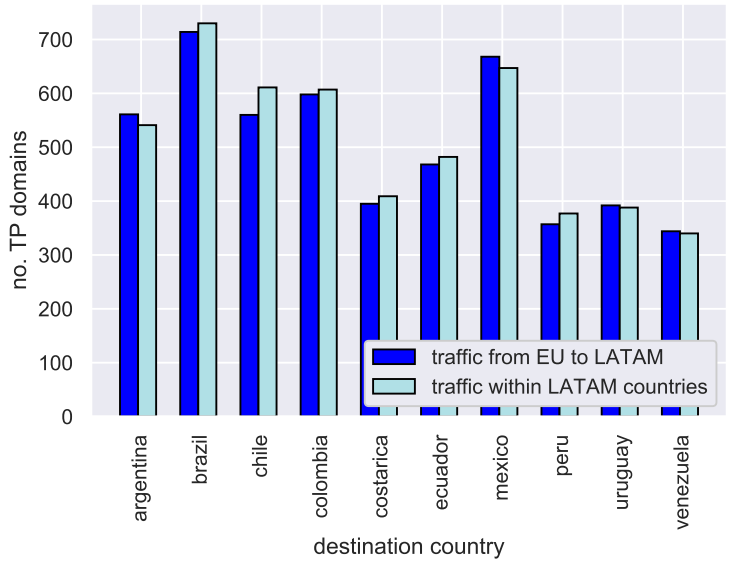


Figure 5.9: Total number of third-party domains found behind web traffic from EU to LATAM.

It is still compelling that very small countries such as Uruguay and Costa Rica might be tracked by a lot of entities despite their small population. When weighting such number of third-party domains by the population of countries, as illustrated in Fig. 5.10, this detail is evidenced. On the other hand, note how huge and highly populated countries such as Brazil and Mexico show an opposite phenomenon: a relatively small number of third-party entities. This again characterizes a marked heterogeneity in the Latin American context, where tiny populations might become the target of several external entities. Naturally, individuals in such groups would be more exposed to privacy risks than those in larger groups.

For some countries, the number of identifying cookies is slightly greater when web requests are generated locally (Argentina, Costa Rica, Ecuador, Peru) than when coming from EU, although in Chile such number is double (see Fig. 5.11). The opposite occurs in the rest of the countries (Brazil, Colombia, Mexico, Uruguay). The inherent potential tracking, then, varies from country to country and apparently regardless its origin is EU or LATAM. Probably, in this regard, a more individualized study should be developed to unveil the reasons of this behavior.

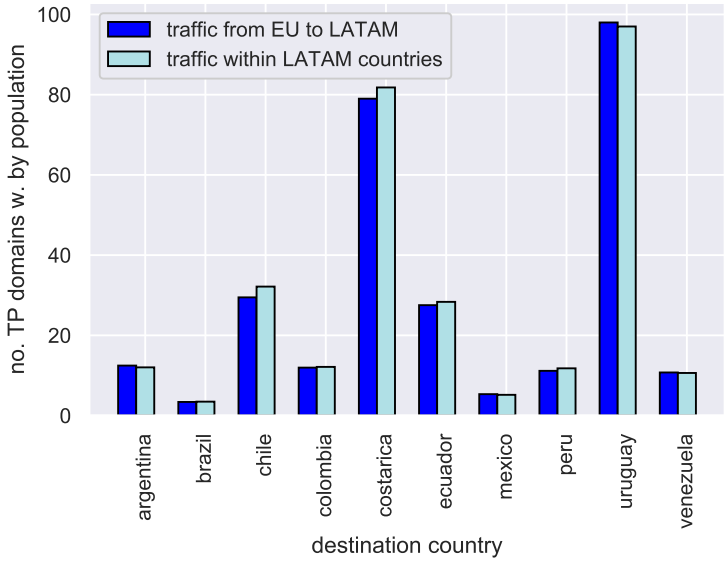


Figure 5.10: Total number of third-party domains found behind web traffic from EU to LATAM.

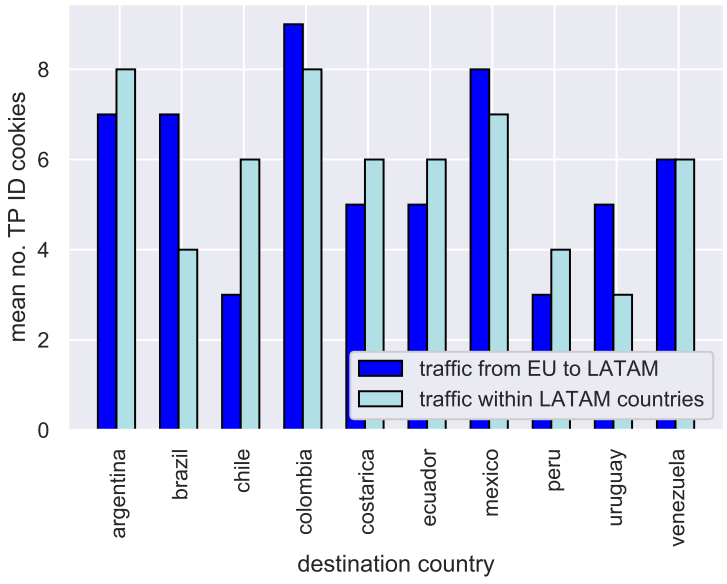


Figure 5.11: Mean number of third-party ID cookies found behind web traffic from EU to LATAM.

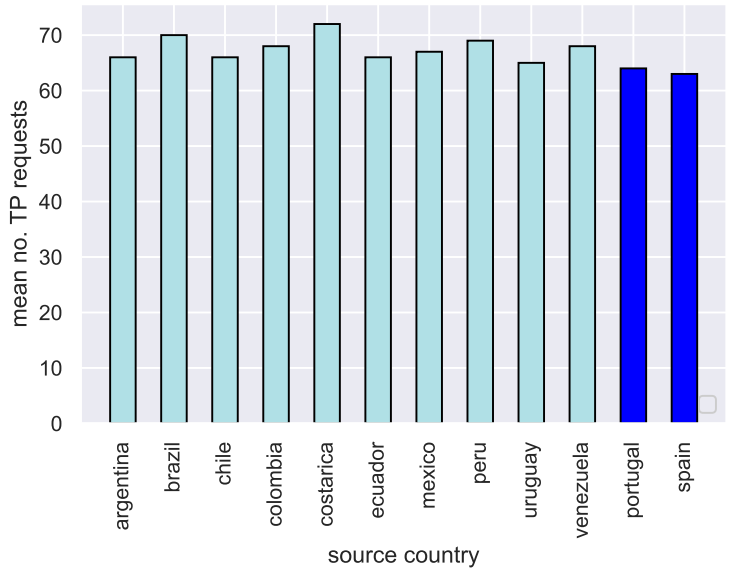


Figure 5.12: Mean number of third-party requests triggered by web traffic to top-world sites.

When testing web traffic from Iberoamerican countries *to the top-world sites*, we found out in Fig. 5.12 that the number of derived third-party requests seems to be, in general, stable along all the countries analyzed. Interestingly, the same effect is shown when illustrating in Fig. 5.13 the total number of third-party domains behind such interactions. Note, however, that this number of entities is almost ten times greater than that when the web traffic was directed to local sites.

The behavior of third-party tracking may be significantly more aggressive when visiting globally popular sites, no matter the source of such visits. Thus, in this scenario, the relative impact on small countries such us Costa Rica, Uruguay and Portugal might be greater given its small population.

The mean number of identifying cookies set when visiting top-world sites from Iberoamerica is also homogeneous along the countries tested. This number is certainly greater than when web traffic is directed to LATAM countries, despite the number of entities behind is ten times higher. That the number of potential tracking cookies does not grow as significantly as the number of third-party domains may suggest that widely popular websites could be also resorting to other more sophisticated tracking mechanisms.

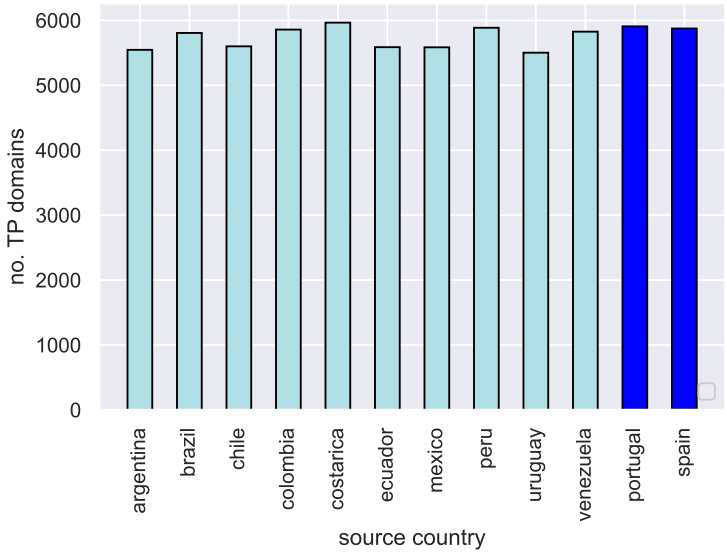


Figure 5.13: Total number of third-party domains found behind web traffic to top-world sites.

The mean number of identifying cookies set when visiting top-world sites from Iberoamerica is also homogeneous along the countries tested. This number is certainly greater than when web traffic is directed to LATAM countries, despite the number of entities behind is ten times higher. That the number of potential tracking cookies does not grow as significantly as the number of third-party domains may suggest that widely popular websites could be also resorting to other more sophisticated tracking mechanisms.

Finally, as shown in Fig. 5.14, `news/media` and `shopping/travel` were again the categories spawning more third-party tracking when browsing top-world sites. Since a wider spectrum of sites was covered in this experiment, other categories were also relevant in terms of potential tracking such as `weather`, `lottery` and `health`. Regarding privacy risks, note how, in this context, sites potentially serving, collecting and even sharing very sensitive information (health or pornography) could be the source of intense third-party tracking.

In any case, gateways of information such as digital newspapers or sites involving any kind of commerce are those spawning a great deal of tracking, either because of their big audiences or because user willingness to buy in these sites is high, no matter where users are located.

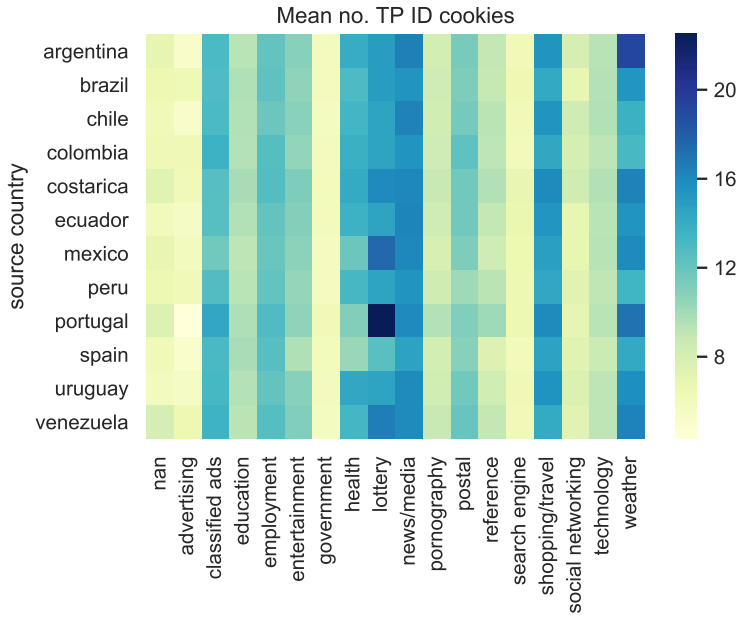


Figure 5.14: Mean number of ID third-party cookies set by web traffic to top-world sites, from Iberoamerican countries, organized by category of publisher.

5.3.2 Online Advertising

Analysis through the `ads.txt` standard

In Fig. 5.15, the adoption of the standard `ads.txt` is depicted. Longer curves such as those of Portugal and Spain indicate that more websites host this file including the ad exchanges authorized to sell ad spaces from said websites. The figure shows that the adoption of `ads.txt` in Latin American countries is still reduced compared to the countries analyzed in the EU. Only Brazil and Argentina present similar levels of adoption. In any case, the number of records per website goes from 1 to 150 for most countries. The specific domains involved are studied below.

To show how third-party entities are distributed along websites in Iberoamerica, based on the records found in the `ads.txt` files, we first plotted Fig. 5.16. This ECDF shows that 1% of the third-party domains found appear on *more* than 20% of the websites crawled (more than 70% of the sites hosting an `ads.txt`). Thus, the concentration of advertising in a few entities is evident in this context; to give an example, `google.com` were engaged with all the sites that had adopted this standard,

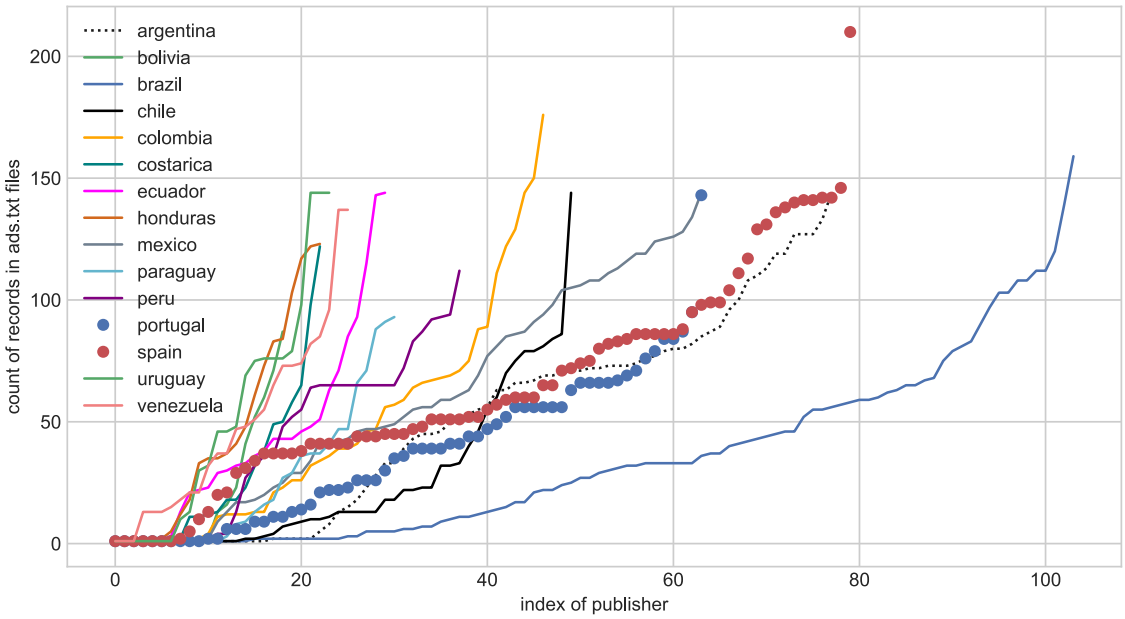


Figure 5.15: Number of records in `ads.txt` files found in websites of Iberoamerican countries.

and 4 others (`rubicon, appnexus, openx and pubmatic`) were engaged with more than 70% of the sites that had an `ads.txt` file.

When analyzing websites by category in Table 5.1, we found that `ads.txt` is widely adopted by `news/media` sites, followed by `entertainment` and `shopping/travel`, as it happened with third-party tracking. Besides, in Table 5.2, we depict the mean number of third-party domains identified within `ads.txt` files. Although, a few `education` related publishers were found that have adopted this standard, the records in the `ads.txt` file in these sites included even more third-party domains than `news/media`. Note that no government hosted an `ads.txt` file in this scenario.

Analysis through the ad related traffic

When analyzing ad related traffic generated in Latin American sites, as a result of browsing from LATAM and EU, we can see that, in general, web traffic from EU triggers a little more ad related tracking than traffic from LATAM, except in Chile and Peru. This was illustrated in Fig. 5.17 where the mean number of ad related requests spawned by country is presented. Websites from Argentina trigger, by far, the highest intensity of ad related traffic, followed by Venezuela. The context

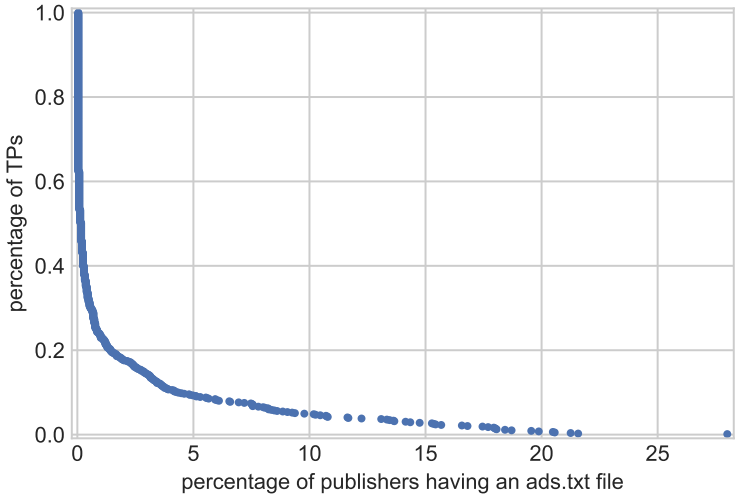


Figure 5.16: ECDF % of web sites covered by third-party domains in Iberoamerican countries.

category	**no. of websites with** `ads.txt`
`news/media`	337
`entertainment`	138
`shopping/travel`	105
`employment`	29
`education`	13
`classified ads`	12
`search engine`	8
`weather`	7
`lottery`	2
`legal`	2
`postal`	1
`health`	1

Table 5.1: Number of websites having adopted the `ads.txt` standard in Iberoamerican countries by category.

is certainly heterogenous but, as stated in previous paragraphs, the source of web traffic might be influencing the ad related tracking derived when browsing LATAM local sites.

We can see that traffic from EU to LATAM would trigger more ad related tracking than traffic to LATAM. Thus, the location of users in this scenario would affect the level of user tracking. Furthermore, if user browsing originates in LATAM and directs

category	mean no. of third-party domains in ads.txt
health	82
education	61.3
news/media	57.69
search engine	49.5
entertainment	47.42
lottery	33.5
classified ads	24.17
employment	22.89
weather	22.85
shopping/travel	15.4
legal	11.5
postal	1

Table 5.2: Mean number of third-party domains found in ads.txt files by category of website in Iberoamerican countries.

to top-world sites, this type of tracking increases notably; its proportion with respect to all third-party traffic is, however, reduced. Consequently, we confirm that users browsing very popular websites would be more exposed to ad related tracking.

Based on the content served by publishers, we found that categories news/media and entertainment provoke the highest ad related traffic both when web traffic comes from LATAM and from EU. For some countries, employment and weather caused important ad related traffic. This is depicted in Figs. 5.18 and 5.19.

Moreover, the impact of this third-party traffic in government sites is minor, but it is present in all countries, although, we feel, commercial advertising for profit should not be engaged with public sites already funded by the taxes of citizens. We found out 103 and 135 government sites triggering ad related tracking from visiting LATAM sites locally and from EU, respectively. More than 90% of these government sites were covered by third-party domains coupled with Google.

When measuring ad related requests triggered by web traffic to top-world publishers, news/media sites showed an important ad tracking activity, followed by entertainment, as with the contexts analyzed previously. This is shown in 5.20. However, in these very popular sites, a significant tracking activity is also observed for several other categories such as employment, education and health. The intrinsic

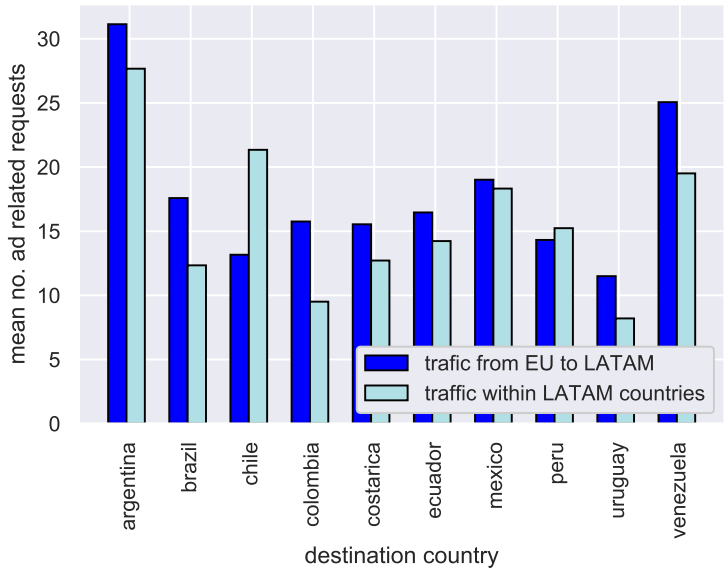


Figure 5.17: Mean number of ad related requests spawned from web traffic to LATAM when originated locally (LATAM) and from EU.

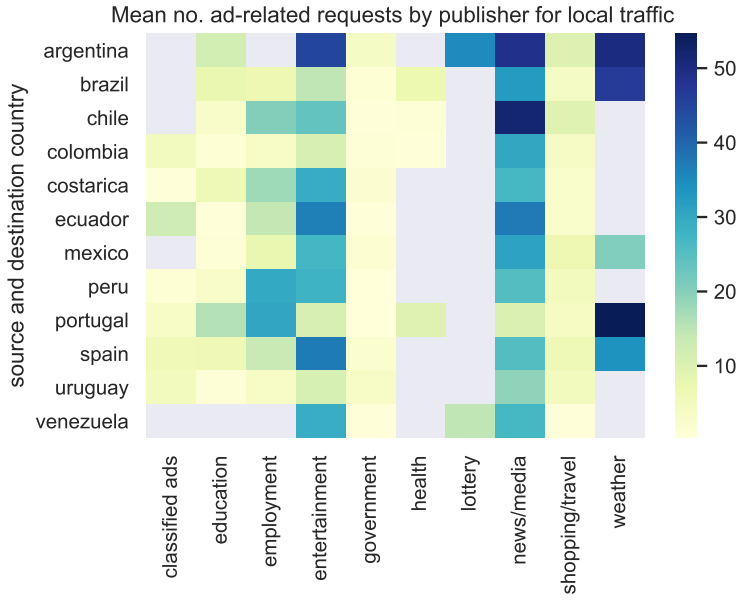


Figure 5.18: Mean number of ad related requests triggered by local traffic.

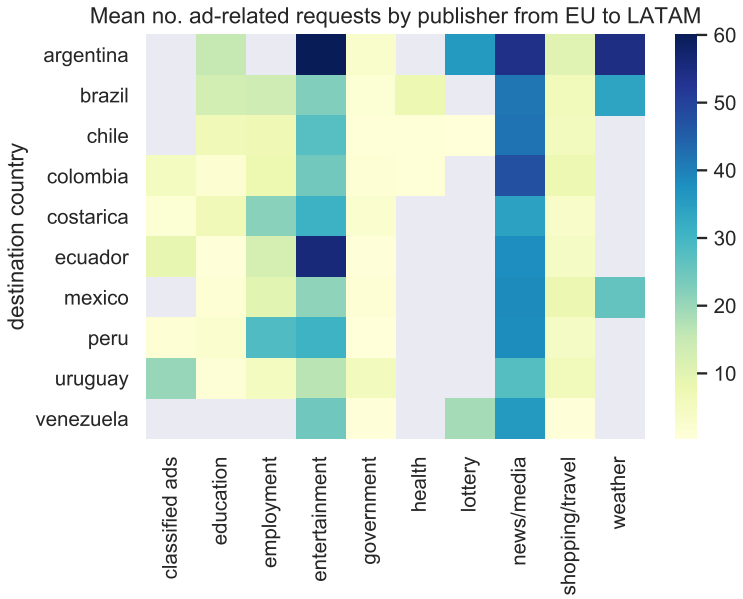


Figure 5.19: Mean number of ad related requests triggered by web traffic from EU to LATAM.

privacy risks involved in this interactions, thus, might be present in a wider spectrum of publishers, including some categorized as `education` and `health` websites, which could be collecting, processing and sharing very sensitive data.

We calculated the prevalence of these third-party entities along local websites in Iberoamerica, and presented the data in Tables 5.3, and 5.4. We found that Google owned domains were massively behind advertising traffic. In particular among the 10 most prevalent third-party domains, at least half of them belonged to this company. `doubleclick.net`, for example, appeared at more than 40% of the web sites crawled from LATAM, a similar coverage of `google.es` when web traffic originated in EU. Besides, the third-party domains involved and their prevalence vary if traffic to LATAM originates in EU; the prevalence of some third-party domains along websites is, in fact, slightly higher. Again, this would suggest that users browsing from EU would be more exposed to the inherent user profiling and tracking of online advertising than LATAM users.

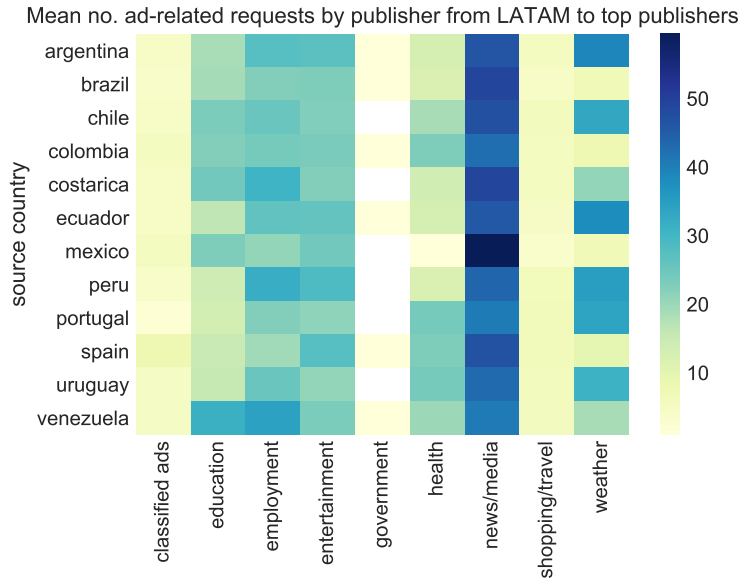


Figure 5.20: Mean number of ad related requests triggered by web traffic from top-world sites.

third-party domain	% of websites covered
doubleclick.net	42.41
googlesyndication.com	31.27
google.com	27.63
ampproject.org	17.54
pubmatic.com	16.38
2mdn.net	12.06
google.com.co	9.00
google.com.br	7.67
yahoo.com	7.38
google.com.ar	5.88

Table 5.3: Prevalence of third-party domains behind ad related traffic along websites visited in Iberoamerican countries *locally*.

5.4 Conclusions

User location affects the intensity of third-party traffic, including advertising related flows of information, triggered from browsing Iberoamerican websites. This is evident first when assessing local web traffic, i.e., when browsing these sites from their country of origin. LATAM and EU countries in this region showed a significant heterogeneity

third-party domain	% of websites covered
google.es	43.27
doubleclick.net	35.71
googlesyndication.com	25.79
google.com	22.67
2mdn.net	15.40
ampproject.org	13.56
pubmatic.com	13.38
yahoo.com	5.42
googleadservices.com	4.09
krxd.net	3.98

Table 5.4: Prevalence of third-party domains behind ad related traffic along websites visited in Iberoamerican countries *from EU*.

among them in this respect. But the influence of location is also shown when measuring the interactions spawned by web traffic from EU to LATAM. We found out that traffic originating in EU spawns more third-party interactions. Despite stricter privacy regulations in EU, users from such locations would be more exposed than users from LATAM to the targeting and profiling performed by external trackers.

Interestingly, the total number of potential trackers (third-party domains) found behind third-party requests, including ad related requests, is similar either web traffic is local or from EU. It seems the same entities are involved in third-party traffic but change their tracking strategy depending on the location of the user.

Users from particular locations might be more "relevant" for online tracking and advertising. Despite the very small population of some countries, the level of third-party traffic derived from them is comparable to that of other huge countries. Portugal is an example.

When top-world sites are the destination of web traffic generated from Iberoamerica, the user location parameter is less relevant to the intensity of potential third-party tracking, since the mean number of requests triggered is similar among the source Iberoamerican countries. Privacy risks arise higher in this context since more third-party requests are spawned and significantly more third-party domains are found behind.

Websites or publishers whose content falls into the categories `news/media`, `entertainment` or `shopping/travel` showed the greatest levels of potential third-party tracking and third-party entities. Thus, the interactions with said websites might entail more privacy risks, as corroborated by previous work in different contexts. Note that when web traffic goes to top-world websites, this risk is extended to other categories related to the collection of sensitive data, such as `health` or `education`.

The level of adoption of the `ads.txt` standard in LATAM is still low compared to that of the EU. The information it reveals is valuable to study the third-parties (ad exchanges) officially coupled with publishers. It evidences the concentration of advertising in a single company, Google; and in publishers associated with the categories `news/media`, `entertainment` or `shopping/travel`.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Online advertising is based on tracking technologies to "follow" and monitor users wherever they browse on the Web. Over time, during such tracking, a lot of metadata is collected about a user, which can be employed to build detailed profiles. Moreover, due to the pervasiveness of online advertising, billions of users get involved in this process. Such range and the powerful personalization technologies considered have made online advertising a millionaire business whose revenues are said to be financing the current free Internet access model.

Since this successful business revolves around the massive exploitation of user data, there are multiple privacy concerns. This is compounded by the complexity and opacity of the internal structure of the online advertising ecosystem, which further complicates the study of privacy risks, and limits the implementation of protection approaches. As a consequence, privacy protection initiatives neglect important parameters, such as the sustainability of the entire system, making them impractical in the long term.

To address this issue, we have performed an exhaustive and global overview of the potential hazards to user privacy in the online advertising ecosystem. An attacker model was built upon the capabilities of main players and the variety of user data likely to be collected. This model encompasses a taxonomy of the roles that potential

attackers might play according, e.g., to their position in the landscape or the amount of user data items they are capable of collecting. Most of the privacy risks arise from ad platforms' intrinsic weaknesses that include non-transparent infrastructure for users, massive collection of user data by thousands intermediary entities, disincentives to incorporate protection strategies, unregulated cooperation among third parties, non-consented collection of granular user data, use of insecure channels to transport sensitive data, and intrusive tracking mechanisms to identify users. In a nutshell, the ecosystem entails a surveillance infrastructure devoted to identify and capture the attention of users. If this where not enough, the participation of users is entirely passive in the ad delivery process, although their interactions fuel the operation of online advertising services.

The consequence of the procedures implemented within this ecosystem is a massive trade of user profiles to perform an abusive identification and classification of users, which could even lead to discrimination. Furthermore, the mechanisms developed to support online tracking and advertising could be so intrusive that persist even after users explicitly delete their fingerprints. Commonly, cookies are used to identify users and track them, and, in combination with other technologies such as canvas, device fingerprinting or cookie matching, to allow cooperation among online advertising entities..

In line with our methodical analysis of privacy risks, we researched available protection approaches. A taxonomy was also construed to organize these contributions. Most of the solutions we have found were based on local mechanisms that block third-party interactions triggered from users to online advertising platforms, thus stopping user data leakage. Other filtering approaches were envisioned as brokers that, interestingly, would allow an active integration of users into the advertising ecosystem. Said integration implies providing more transparency and control for users, even though blocking third-party tracking might be affecting the economic model of the Web. Advertising schemes offering native privacy have been proposed but unfortunately involve important modifications of the current structure, thus complicating their implementation.

With the aim of scrutinizing the privacy risks arising within the context of ad platforms, we take advantage of data obtained from bid interactions released by a known Chinese DSP. In particular, messages sent by an ad platform to DSPs, inviting them to participate in auctions, contain granular data about users' behavior. Since this data might be distributed, without regulation, among hundreds of such DSPs, they become potential massive surveillance entities, even at a very low cost. This is partially due to the power given to advertisers to target users with very high precision.

In order to mitigate this problem, we have proposed reducing the amount of DSPs that receive bid requests while preserving the revenue of the system. Unlike radical blocking approaches, probably impractical in the long term, we have looked for a balanced trade-off between the number of invited DSPs (trying to preserve privacy) and revenue. An optimal solution is mathematically viable that maximizes revenue while bounding the participation of DSPs. When applying it, though revenue certainly increases with the number of invited DSPs, an important portion of revenue is still preserved even a significant reduction in the DSPs is enforced. Interestingly, this strategy, i.e., penalizing abusing DSPs, could encourage them to "behave" better.

[Conclusions last paper]

6.2 Future Work

The complexity of online advertising poses various challenges to user privacy. From the analysis conducted in the previous sections, we envisage two main lines of future research: identifying new privacy threats and providing new protection mechanisms. Given the opacity of ad platforms, we believe that further exploring the tracking capabilities of the advertising industry will help discover their potential to become privacy attackers. But not only that, unveiling the user data exchange processes within ad platforms would expose the extent to which some of their intermediate entities are prone to become massive surveillance agents. A better knowledge of the adversary will contribute to develop protection mechanisms which are more tailored to the above mentioned privacy threats.

As for privacy mechanisms, a natural next step would be combining some of the proposals described in Sec. 3.3. Such synergy shall generate more robust and useful privacy solutions for detecting user-related flaws and invasive tracking behaviors, and better adapting privacy enhancing technologies to the current Web economic model.

Regarding the strategy posed by current privacy protection approaches (namely blocking, obfuscation, sandboxing, and user inclusion), a further analysis on their impact on the Web economic model will reveal if such proposals are effectively adapting to the current advertising business model, without a significant side effect.

A further research direction for improving users' privacy in online advertising is to create smarter protection tools in the user side, that is, developed as browser complements. Intelligibility, usability and flexibility are some of the parameters that need to be considered to enable mechanisms to give users real transparency and control over their browsing data. In this regard, a great deal of work has to be done to develop tools that let users effectively enforce their motivations on the protection strategy selected.

Another strand of research may consider the scope of the protection strategy, currently limited to the user side. Extending the scope of the privacy protection mechanisms to the different players (e.g., publishers, advertisers, ad exchanges) might result in a more solid approach. Accordingly, analyzing and evaluating the privacy policies and protection mechanisms offered by ad platforms might contribute to detect their flaws and make improvements.

To go beyond the simplistic (and endangering) blocking strategy of some approaches examined in Sec. 3.3, new advertising models have to be envisioned that provide flexible two-way communication interfaces between users and ad platforms through which they could directly manage their relationship according to their interests. While economic interest of advertising entities are widely known, user motivations related to privacy, advertising choices and even economic incentives should be seriously considered by such models. Undoubtedly, more transparent and balanced interactions will derive in an increased sense of security and thus of privacy.

A more user-driven advertising platform, where user interests regarding their privacy and profit may be variable (not always opposing to the advertisers'), and the

assessment of user information as an asset with intrinsic economic value, not only for intermediate advertising entities, but also for users, will help to study the trade-off between such value and the privacy of users involved in online advertising transactions.

Furthermore, the ad delivery model itself must be rethinked because its components can implement privacy more effectively, in particular, concerning powerful privacy techniques such as data minimization and transparency for users.

Bibliography

[1] Disconnect.me, "Private browsing," accessed on 2015-12-16. [Online]. Available: https://disconnect.me/disconnect

[2] M. Smith, *Targeted: How Technology Is Revolutionizing Advertising and the Way Companies Reach Consumers*, 1st ed. New York: AMACOM, Nov. 2014.

[3] "Real-time bidding protocol - cookie matching," accessed on 2015-10-07. [Online]. Available: https://developers.google.com/ad-exchange/rtb/cookie-guide

[4] S. Yuan, A. Z. Abidin, M. Sloan, and J. Wang, "Internet advertising: An interplay among advertisers, online publishers, ad exchanges and web users," *arXiv: 1206.1754*, 2012, arXiv preprint.

[5] D. Lyon, Ed., *Surveillance as Social Sorting: Privacy, Risk and Automated Discrimination*. Routledge, Dec. 2002.

[6] U. Senate, "Online advertising and hidden hazards to consumer security," Tech. Rep., 2014. [Online]. Available: http://www.hsgac.senate.gov/hearings/online-advertising-and-hidden-hazards-to-consumer-security-and-data-privacy

[7] D. S. Evans, "The online advertising industry: Economics, evolution, and privacy," *The journal of economic perspectives*, vol. 23, no. 3, pp. 37–60, 2009.

[8] "The cost of ad blocking," PageFair, Res. Rep., Aug. 2015.

[9] "The state of online advertising," Adobe, Tech. Rep., 2012, accessed on 2015-09-11. [Online]. Available: http://www.adobe.com/aboutadobe/pressroom/pdfs/Adobe_State_of_Online_Advertising_Study.pdf

[10] G. Marvin, "Consumers now notice retargeted ads," Marketing Land, Tech. Rep., Dec. 2013, accessed on 2015-08-12. [Online]. Available: http://marketingland.com/3-out-4-consumers-notice-retargeted-ads-67813

[11] M. Smith, *Targeted: How technology is revolutionizing advertising and the way companies reach consumers.* AMACOM Div American Mgmt Assn, 2014.

[12] "Real-time bidding protocol - processing the request," accessed on 2017-04-07. [Online]. Available: https://developers.google.com/ad-exchange/rtb/request-guide

[13] H. Nissenbaum, *Privacy in context: Technology, policy, and the integrity of social life.* Stanford University Press, 2009.

[14] L. Olejnik, T. Minh-Dung, and C. Castelluccia, "Selling off privacy at auction," in *Proc. Symp. Netw. Distrib. Syst. Secur. (SNDSS)*, Feb. 2014.

[15] T. Speicher, M. Ali, G. Venkatadri, F. N. Ribeiro, G. Arvanitakis, F. Benevenuto, K. P. Gummadi, P. Loiseau, and A. Mislove, "Potential for discrimination in online targeted advertising," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 5–19.

[16] J. Estrada-Jiménez, J. Parra-Arnau, A. Rodríguez-Hoyos, and J. Forné, "On the regulation of personal data distribution in online advertising platforms," *Engineering Applications of Artificial Intelligence*, vol. 82, pp. 13–29, 2019.

[17] J. Estrada-Jiménez, J. Parra-Arnau, A. Rodríguez-Hoyos, and J. Forné, "Online advertising: Analysis of privacy threats and protection approaches," *Computer Communications*, vol. 100, pp. 32–51, 2017.

[18] J. Estrada-Jiménez, A. Rodríguez-Hoyos, J. Parra-Arnau, and J. Forné, "Measuring online tracking and privacy risks on ecuadorian websites," in *2019 IEEE Fourth Ecuador Technical Chapters Meeting (ETCM)*. IEEE, 2019, pp. 1–6.

[19] A. Rodríguez-Hoyos, D. Rebollo-Monedero, J. Estrada-Jiménez, and J. Forné.

[20] E. Pallarès, D. Rebollo-Monedero, A. Rodríguez-Hoyos, J. Estrada-Jiménez, A. Mohamad Mezher, and J. Forné, "Mathematically optimized, recursive prepartitioning strategies for k-anonymous microaggregation of large-scale datasets," vol. 144.

[21] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, A. M. Mezher, J. Parra-Arnau, and J. Forné, "The fast MDAV (F-MDAV) algorithm: An algorithm for k-anonymous microaggregation in big data," 2020.

[22] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, J. Forné, R. Trapero, A. Álvarez, and R. Rodríguez, "Anonymizing cybersecurity data in critical infrastructures: The cipsec approach," in *2019 International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. ISCRAM, 2019.

[23] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, "Does k-anonymous microaggregation affect machine-learned macrotrends?" vol. 6, pp. 28 258–28 277, May 2018. [Online]. Available: http://doi.org/10.1109/ACCESS.2018.2834858

[24] A. Rodríguez-Hoyos, J. Estrada-Jiménez, L. Urquiza-Aguiar, J. Parra-Arnau, and J. Forné, "Digital hyper-transparency: leading e-government against privacy," in *2018 International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, 2018, pp. 263–268.

[25] S. Yuan, J. Wang, and X. Zhao, "Real-time bidding for online advertising: measurement and analysis," in *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*. ACM, 2013, p. 3.

[26] H. Beales, "The value of behavioral targeting," Netw. Advertising Initiative, Tech. Rep., Mar. 2010, accessed on 2016-01-15. [Online]. Available: http://www.networkadvertising.org/pdfs/Beales_NAI_Study.pdf

[27] A. M. McDonald and L. F. Cranor, "Americans' attitudes about internet behavioral advertising practices," in *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*. ACM, 2010, pp. 63–72.

[28] F. Rejón-Guardia and F. J. Martínez-López, "Online advertising intrusiveness and consumers' avoidance behaviors," in *Handbook of Strategic e-Business Management*. Springer, 2014, pp. 565–586.

[29] M. Schudson, *Advertising, the uneasy persuasion (RLE Advertising): Its dubious impact on American society*. Routledge, 2013.

[30] D. Morgan, "TV audience fragmentation is an inescapable reality: Embrace it," Sep. 2012, accessed on 2016-03-04. [Online]. Available: http://www.mediapost.com/publications/article/182946/tv-audience-fragmentation-is-an-inescapable-realit.html

[31] K. Nelson-Field and E. Riebe, "The impact of media fragmentation on audience targeting: An empirical generalisation approach," *Journal of Marketing Communications*, vol. 17, no. 01, pp. 51–67, 2011.

[32] K. McSpadden, "You now have a shorter attention span than a goldfish," May 2015, accessed on 2016-03-04. [Online]. Available: http://time.com/3858309/attention-spans-goldfish/

[33] P. Minnium, "8 reasons why digital advertising works for brands," Nov. 2014, accessed on 2016-03-12. [Online]. Available: http://marketingland.com/10-reasons-digital-advertising-works-brands-108151

[34] A. A. Kardan and M. Hooman, "Targeted advertisement in social networks using recommender systems," in *e-Commerce in Developing Countries: With*

Focus on e-Security (ECDC), 2013 7th Intenational Conference on. IEEE, 2013, pp. 1–13.

[35] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen, "How much can behavioral targeting help online advertising?" in *Proceedings of the 18th international conference on World wide web.* ACM, 2009, pp. 261–270.

[36] C. E. Tucker, "Social networks, personalized advertising, and privacy controls," *Journal of Marketing Research*, vol. 51, no. 5, pp. 546–562, 2014.

[37] C. Gayomali, "It would cost each user $232 a year for an ad-free internet, study finds," Aug. 2014, accessed on 2016-02-27. [Online]. Available: http://www.fastcompany.com/3034670/fast-feed/it-would-cost-each-user-232-a-year-for-an-ad-free-internet-study-finds

[38] IAB, "Digital ad revenues surge 19 percent, climbing to $27.5 billion in first half of 2015," Oct. 2015, accessed on 2016-03-06. [Online]. Available: http://www.iab.com/news/digital-ad-revenues-surge-19-climbing-to-27-5-billion-in-first-half-of-2015-according-to-iab-internet-

[39] OpenX, "Ad networks vs. ad exchanges: How they stack up," Jul. 2010, accessed on 2016-03-06. [Online]. Available: http://openx.com/blog/openx-releases-new-whitepaper-ad-networks-vs-ad-exchanges/

[40] J. R. Mayer and J. C. Mitchell, "Third-party web tracking: Policy and technology," in *2012 IEEE Symposium on Security and Privacy.* IEEE, 2012, pp. 413–427.

[41] Google, "Cookie matching," Mar. 2016, accessed on 2016-03-07. [Online]. Available: https://developers.google.com/ad-exchange/rtb/cookie-guide#examples

[42] D. Laffey, "Paid search: The innovation that changed the web," *Business Horizons*, vol. 50, no. 3, pp. 211–218, 2007.

[43] S. Horowitz, "Real time bidding - rtb - programmatic advertising explained," accessed on 2016-03-09. [Online]. Available: http://executive-digital.com/digital-marketing-experts/what-is-real-time-bidding-and-why-is-it-more-effective-than-direct-bidding-methods/

[44] A. Ghosh, M. Mahdian, R. P. McAfee, and S. Vassilvitskii, "To match or not to match: Economics of cookie matching in online advertising," *ACM Transactions on Economics and Computation*, vol. 3, no. 2, p. 12, 2015.

[45] L. Olejnik and C. Castelluccia, "To bid or not to bid? measuring the value of privacy in RTB," accessed on 2016-05-21. [Online]. Available: http://lukaszolejnik.com/rtb2.pdf

[46] P. Eckersley, "How unique is your web browser?" in *Privacy Enhancing Technologies*, vol. 6205. Springer, 2010, pp. 1–18.

[47] "Adblock Plus - surf the Web without annoying ads!" Nov. 2015, accessed on 2015-11-15. [Online]. Available: https://adblockplus.org

[48] J. Parra-Arnau, J. P. Achara, and C. Castelluccia, "Myadchoices: Bringing transparency and control to online advertising," *arXiv preprint arXiv:1602.02046*, 2016.

[49] D. Sánchez and A. Viejo, "Privacy-preserving and advertising-friendly web surfing," *Computer Communications*, vol. 130, pp. 113–123, 2018.

[50] J. P. Achara, J. Parra-Arnau, and C. Castelluccia, "Mytrackingchoices: Pacifying the ad-block war by enforcing user privacy preferences," *arXiv preprint arXiv:1604.04495*, 2016.

[51] M. Backes, A. Kate, M. Maffei, and K. Pecina, "Obliviad: Provably secure and practical online behavioral advertising," in *2012 IEEE Symposium on Security and Privacy*. IEEE, 2012, pp. 257–271.

[52] S. Guha, B. Cheng, and P. Francis, "Privad: practical privacy in online advertising," in *USENIX conference on Networked systems design and implementation*, 2011, pp. 169–182.

[53] Privoxy.org, "Privoxy," Mar. 2016, accessed on 2016-03-17. [Online]. Available: http://www.privoxy.org

[54] "Easylist - overview," Mar. 2016, accessed on 2016-05-30. [Online]. Available: https://easylist.github.io

[55] P. Papadopoulos, N. Kourtellis, and E. P. Markatos, "Cookie synchronization: Everything you always wanted to know but were afraid to ask," *arXiv preprint arXiv:1805.10505*, 2018.

[56] M. A. Bashir, S. Arshad, E. Kirda, W. Robertson, and C. Wilson, "How tracking companies circumvented ad blockers using websockets," in *Proceedings of the Internet Measurement Conference 2018*. ACM, 2018, pp. 471–477.

[57] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas, "Adnostic: Privacy preserving targeted advertising," in *Proc. Symp. Netw. Distrib. Syst. Secur. (SNDSS)*, Feb. 2010, pp. 1–21.

[58] M. Fredrikson and B. Livshits, "Repriv: Re-envisioning in-browser privacy," in *Proc. IEEE Symp. Security, Privacy (SP)(May 2011)*, 2010.

[59] L. J. Helsloot, G. Tillem, and Z. Erkin, "Badass: Preserving privacy in behavioural advertising with applied secret sharing," in *International Conference on Provable Security*. Springer, 2018, pp. 397–405.

[60] L. J. Helsloot, G. Tillem, and Z. Erkin, "Ahead: Privacy-preserving online behavioural advertising using homomorphic encryption," in *Information Forensics and Security (WIFS), 2017 IEEE Workshop on*. IEEE, 2017, pp. 1–6.

[61] J. Parra-Arnau, "Pay-per-tracking: A collaborative masking model for web browsing," *Information Sciences*, vol. 385, pp. 96–124, 2017.

[62] Brave, "Brave software," Mar. 2016, accessed on 2016-05-30. [Online]. Available: https://www.brave.com

[63] Mozilla, "Subscribe2web," accessed on 2015-11-21. [Online]. Available: https://air.mozilla.org/subscribe2web/

[64] B. Shiller, J. Waldfogel, and J. Ryan, "Will ad blocking break the internet?" National Bureau of Economic Research, Tech. Rep., 2017.

[65] E. GDPR, "Home page of eu gdpr," *línea]. Disponible en: https://www. eugdpr. org/.[Accedido: 19-feb-2018]*.

[66] I. Faizullabhoy and A. Korolova, "Facebook's advertising platform: New attack vectors and the need for interventions," *arXiv preprint arXiv:1803.10099*, 2018.

[67] G. Venkatadri, A. Andreou, Y. Liu, A. Mislove, K. P. Gummadi, P. Loiseau, and O. Goga, "Privacy risks with facebook's pii-based targeting: Auditing a data broker's advertising interface," in *IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 221–239.

[68] A. Goldfarb and C. E. Tucker, "Privacy regulation and online advertising," *Management science*, vol. 57, no. 1, pp. 57–71, 2011.

[69] B. Ur, P. G. Leon, L. F. Cranor, R. Shay, and Y. Wang, "Smart, useful, scary, creepy: perceptions of online behavioral advertising," in *proceedings of the eighth symposium on usable privacy and security*. ACM, 2012, p. 4.

[70] "US programmatic ad spend tops $10 billion this year, to double by 2016," eMarketer, Tech. Rep., Oct. 2014. [Online]. Available: http://www.emarketer.com/Article/US-Programmatic-Ad-Spend-Tops-10-Billion-This-Year-Double-by-2016/1011312

[71] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Measuring the privacy of user profiles in personalized information systems," *Future Generation Computer Systems*, vol. 33, pp. 53–63, 2014.

[72] S. Pandey, M. Aly, A. Bagherjeiran, A. Hatch, P. Ciccolo, A. Ratnaparkhi, and M. Zinkevich, "Learning to target: What works for behavioral targeting," in *Proc. Int. Conf. Inform., Knowl. Manage. (CIKM)*. ACM, 2011, pp. 1805–1814.

[73] "How real a threat is de-anonymization?" May 2011, accessed on 2016-03-09. [Online]. Available: https://swildstrom.wordpress.com/2011/05/31/how-real-a-threat-is-de-anonymization/

[74] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur., Priv. (SP)*. IEEE Comput. Soc., May 2008, pp. 111–125.

[75] C. A. Dwyer, "Behavioral targeting: A case study of consumer tracking on levis.com," *Available at SSRN 1508496*, 2009.

[76] M. A. Bashir, S. Arshad, W. Robertson, and C. Wilson, "Tracing information flows between ad exchanges using retargeted ads," in *Proceedings of the 25th USENIX Security Symposium*, 2016.

[77] F. Roesner, C. Rovillos, T. Kohno, and D. Wetherall, "Balancing privacy and functionality of third-party social widgets," *USENIX Magazine*, 2012.

[78] N. Vratonjic, M. H. Manshaei, J. Grossklags, and J.-P. Hubaux, "Ad-blocking games: Monetizing online content under the threat of ad avoidance," in *The Economics of Information Security and Privacy*. Springer, 2013, pp. 49–73.

[79] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz., "The web never forgets: Persistent tracking mechanisms in the wild," in *Proc. ACM Conf. Comput., Commun. Secur. (CCS)*, Washington, DC, Nov. 2014, pp. 674–689.

[80] S. Englehardt, D. Reisman, C. Eubank, P. Zimmerman, J. Mayer, A. Narayanan, and E. W. Felten, "Cookies that give you away: The surveillance implications of web tracking," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 289–299.

[81] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 1388–1401.

[82] A. Soltani, S. Canty, Q. Mayo, L. Thomas, and C. J. Hoofnagle, "Flash cookies and privacy," in *Proc. AAAI Spring Symp. Intell. Inform. Priv. Manage.* Assoc. Adv. Artif. Intell., 2010.

[83] M. D. Ayenson, D. J. Wambach, A. Soltani, N. Good, and C. J. Hoofnagle, "Flash cookies and privacy II: Now with HTML5 and ETag respawning," *Available at SSRN 1898390*, 2011.

[84] K. O. Advertising, "Flash cookies, HTML 5 storage and HTTP cookies - know online advertising," accessed on 2016-06-15. [Online]. Available: http://www.knowonlineadvertising.com/difference-between/flash-cookies-html-5-storage-and-http-cookies

[85] A. M. McDonald and L. F. Cranor, "Survey of the use of adobe flash local shared objects to respawn HTTP cookies," *ISJLP*, vol. 7, p. 639, 2011.

[86] Addthis.com, "Get more like, shares and follows with smart website tools," Mar. 2016, accessed on 2016-05-30. [Online]. Available: http://www.addthis.com

[87] C. J. Hoofnagle, A. Soltani, N. Good, D. J. Wambach, and M. D. Ayenson, "Behavioral advertising: the offer you cannot refuse," 2012.

[88] D. Kravets, "Lawsuit targets mobile advertiser over sneaky html5 pseudo-cookies," Sep. 2010, accessed on 2016-05-30. [Online]. Available: https://www.wired.com/2010/09/html5-safari-exploit

[89] M. S. Ackerman, L. F. Cranor, and J. Reagle, "Privacy in e-commerce: examining user scenarios and privacy preferences," in *Proceedings of the 1st ACM conference on Electronic commerce*. ACM, 1999, pp. 1–8.

[90] L. Agarwal, N. Shrivastava, S. Jaiswal, and S. Panjwani, "Do not embarrass: re-examining user concerns for online tracking and advertising," in *Proceedings of the Ninth Symposium on Usable Privacy and Security*. ACM, 2013, p. 8.

[91] J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira, "Your browsing behavior for a big mac: Economics of personal information online," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 189–200.

[92] A. E. Schlosser, S. Shavitt, and A. Kanfer, "Survey of internet users? attitudes toward internet advertising," *Journal of interactive marketing*, vol. 13, no. 3, pp. 34–54, 1999.

[93] L. Sweeney, "Discrimination in online ad delivery," *Queue*, vol. 11, no. 3, p. 10, 2013.

[94] Electronic Frontier Foundation, "Privacy badger," Mar. 2016, accessed on 2016-03-15. [Online]. Available: https://www.eff.org/privacybadger

[95] J. Parra-Arnau, J. P. Achara, and C. Castelluccia, "MyAdChoices: Bringing transparency and control to online advertising," *ACM Trans. Web*, arXiv preprint. [Online]. Available: http://arxiv.org/abs/1602.02046

[96] D. C. Howe and H. Nissenbaum, *Lessons from the Identity Trail: Privacy, Anonymity and Identity in a Networked Society*. NY: Oxford Univ. Press, 2009, ch. TrackMeNot: Resisting surveillance in Web search, pp. 417–436. [Online]. Available: http://mrl.nyu.edu/~dhowe/trackmenot

[97] J. P. Achara, J. Parra-Arnau, and C. Castelluccia, "MyTrackingChoices: Pacifying the ad-block war by enforcing user privacy preferences," in *Proc. Annual Workshop Econ. Inform. Secur. (WEIS)*, Jul. 2016, to appear.

[98] M. Ter Louw, K. T. Ganesh, and V. Venkatakrishnan, "Adjail: Practical enforcement of confidentiality and integrity policies on web advertisements." in *USENIX Security Symposium*, 2010, pp. 371–388.

[99] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas, "Adnostic: Privacy preserving targeted advertising," in *Proceedings Network and Distributed System Symposium*, 2010.

[100] F. Roesner, T. Kohno, and D. Wetherall, "Detecting and defending against third-party tracking on the web," in *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 12–12. [Online]. Available: http://dl.acm.org/citation.cfm?id=2228298.2228315

[101] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan, "Adreveal: Improving transparency into online targeted advertising," in *Proc. Hot Topics in Netw.* ACM, 2013, pp. 12:1–12:7.

[102] "AOL search data scandal," Aug. 2006, accessed on 2013-11-15. [Online]. Available: http://en.wikipedia.org/wiki/AOL_search_data_scandal

[103] M. Lecuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu, "XRay: Enhancing the web's transparency with differential correlation," in *Proc. Conf. USENIX Secur. Symp.*, Aug. 2014.

[104] Ghostery, "How does ghostery make money from the browser extension?" accessed on 2015-12-21. [Online]. Available: https://www.ghostery.com/support/faq/ghostery-add-on/how-does-ghostery-make-money-from-the-add-on/

[105] BetaNews, "Adblock plus updates acceptable ads and reveals how it makes money," 2015, accessed on 2015-12-21. [Online]. Available: http://betanews.com/2015/12/16/adblock-plus-updates-acceptable-ads-and-reveals-how-it-makes-money/

[106] "AdBlock," Mar. 2016, accessed on 2016-03-13. [Online]. Available: https://getadblock.com/

[107] Mozilla, "Lightbeam for firefox," accessed on 2015-12-16. [Online]. Available: https://www.mozilla.org/en-US/lightbeam/

[108] Abine, "Blur: Keep your web activity and personal info private," Mar. 2016, accessed on 2016-03-15. [Online]. Available: https://dnt.abine.com/#dashboard

[109] "SuperBlock Adblocker," accessed on 2016-05-12. [Online]. Available: https://chrome.google.com/webstore/detail/superblock-adblocker/miijbmhjndcihicbljlcieiajhemmdeb

[110] "AdRemover," accessed on 2016-05-11. [Online]. Available: https://chrome.google.com/webstore/detail/adremover-for-google-chro/mcefmojpghnaceadnghednjhbmphipkb

[111] "Adblock Pro," accessed on 2016-05-11. [Online]. Available: https://chrome.google.com/webstore/detail/adblock-pro/ocifcklkibdehekfnmflempfgjhbedch

[112] "uBlock," accessed on 2016-05-11. [Online]. Available: https://www.ublock.org/

[113] PageFair, "The 2015 ad blocking report," 2015, accessed on 2015-11-20. [Online]. Available: https://blog.pagefair.com/2015/ad-blocking-report/

[114] "Allowing acceptable ads in adblock plus," Mar. 2016, accessed on 2016-03-13. [Online]. Available: https://adblockplus.org/en/acceptable-ads

[115] R. Cookson, "Google, Microsoft and Amazon pay to get around ad blocking tool," Feb. 2015, accessed on 2014-03-10. [Online]. Available: http://www.ft.com/cms/s/0/80a8ce54-a61d-11e4-9bd3-00144feab7de.html

[116] M. Sullivan, "Adblock plus was in NYC last week waving the olive branch at advertisers," Nov. 2015, accessed on 2015-11-20. [Online]. Available: http://venturebeat.com/2015/11/13/adblock-plus-was-in-nyc-last-week-waving-the-olive-branch-at-advertisers

[117] "Ghostery - take control of your digital experience," Nov. 2015, accessed on 2015-11-15. [Online]. Available: https://www.ghostery.com/

[118] P. Leon, B. Ur, R. Shay, Y. Wang, R. Balebako, and L. Cranor, "Why Johnny can't opt out: a usability evaluation of tools to limit online behavioral advertising," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 589–598.

[119] A. Henry, "Ad-blocker ghostery actually helps advertisers, if you."

[120] "GoogleSharing." [Online]. Available: www.googlesharing.net

[121] T. Adams, "Can the Web save the press from oblivion?" Apr. 2016, accessed on 2016-05-30. [Online]. Available: http://www.theguardian.com/media/2016/apr/17/can-internet-save-printed-press-blendle-lumi

[122] Google, "Google contributor," Mar. 2016, accessed on 2016-03-15. [Online]. Available: https://www.google.com/contributor/welcome

[123] N. Lomas, "Shine signs first european carriers to its network-level ad blocking tech," Feb. 2016, accessed on 2016-03-17. [Online]. Available: http://techcrunch.com/2016/02/18/shine-bags-first-european-carrier-as-three-uk-deploys-network-level-ad-blocking

[124] N. Lomas, "U.K. carrier EE looking at giving users control over mobile ads," Nov. 2015, accessed on 2016-03-17. [Online]. Available: http://techcrunch.com/2015/11/23/u-k-carrier-ee-looking-at-giving-users-control-over-mobile-ads

[125] A. Acquisti, C. R. Taylor, and L. Wagman, "The economics of privacy," *Available at SSRN 2580411*, 2016.

[126] D. Bergemann and A. Bonatti, "Selling cookies," *American Economic Journal: Microeconomics*, vol. 7, no. 3, pp. 259–294, 2015.

[127] A. De Corniere, "Search advertising," *Available at SSRN 1967102*, 2013.

[128] C. Taylor and L. Wagman, "Consumer privacy in oligopolistic markets: Winners, losers, and welfare," *International Journal of Industrial Organization*, vol. 34, pp. 80–84, 2014.

[129] J. Turow, J. King, C. J. Hoofnagle, A. Bleakley, and M. Hennessy, "Americans reject tailored advertising and three activities that enable it," *Available at SSRN 1478214*, 2009.

[130] L. Rainie, S. Kiesler, R. Kang, M. Madden, M. Duggan, S. Brown, and L. Dabbish, "Anonymity, privacy, and security online," *Pew Research Center*, vol. 5, 2013.

[131] A. W. Sile, "Privacy compromised? might as well monetize," Jan. 2015.

[132] "The age of digital enlightenment," Logicalis, Tech. Rep., Mar. 2016, accessed on 2016-05-17. [Online]. Available: http://www.uk.logicalis.com/globalassets/united-kingdom/microsites/real-time-generation/realtime-generation-2016-report.pdf

[133] "Privacy and security in a connected life: A study of us, european and japanese consumers," Ponemon Institute, Tech. Rep., Mar. 2015, accessed on 2016-05-14. [Online]. Available: http://www.trendmicro.com/vinfo/us/security/news/internet-of-things/internet-of-things-connected-life-security

[134] A. M. Mcdonald, R. W. Reeder, P. G. Kelley, and L. F. Cranor, "A comparative study of online privacy policies and formats," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*. Seattle, WA: Springer-Verlag, 2009, pp. 37–55.

[135] M. Hoelzel, "The programatic-advertising report: Mobile, video, and real-time bidding drive growth in programmatic," Apr. 2015, accessed on 2017-06-30. [Online]. Available: http://www.businessinsider.com/buyers-and-sellers-have-overwhelmingly-adopted-programmatic-with-mobile-leading-growth-2015-4

[136] Google, "Ad exchange auction model," Oct. 2018. [Online]. Available: https://support.google.com/authorizedbuyers/answer/6077702?hl=en

[137] Google, "Real-time bidding protocol," 2017, accessed on 2017-06-02. [Online]. Available: https://developers.google.com/ad-exchange/rtb/downloads/realtime-bidding-proto.txt

[138] L. Sweeney, "k-Anonymity: A model for protecting privacy," *Int. J. Uncertain., Fuzz., Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[139] J. Mathai, G. Ramasamy, S. Purusothaman, and K. Ezra, "Location based mobile advertising framework for commuters," in *Computing and Network Communications (CoCoNet), 2015 International Conference on*. IEEE, 2015, pp. 928–935.

[140] J. Brookman, P. Rouge, A. Alva, and C. Yeung, "Cross-device tracking: Measurement and disclosures," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 2, pp. 133–148, 2017.

[141] K. Hill, "How target figured out a teen girl was pregnant before her father did," Feb. 2012. [Online]. Available: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#764bfe476668

[142] W. Zhang, S. Yuan, J. Wang, and X. Shen, "Real-time bidding benchmarking with ipinyou dataset," *arXiv preprint arXiv:1407.7073*, 2014.

[143] M. N. del Prado Cortez and J. Frignal, "Geo-location inference attacks: From modelling to privacy risk assessment (short paper)," in *Dependable Computing Conference (EDCC), 2014 Tenth European*. IEEE, 2014, pp. 222–225.

[144] Google, "Google doubleclick ad exchange (adx) buyer program guidelines," Jun. 2017. [Online]. Available: https://www.google.com/doubleclick/adxbuyer/guidelines.html

[145] A. Korolova, "Privacy violations using microtargeted ads: A case study," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*. IEEE, 2010, pp. 474–482.

[146] B. Collins, "Russia's facebook fake news could have reached 70 million americans," Aug. 2017. [Online]. Available: http://www.thedailybeast.com/russias-facebook-fake-news-could-have-reached-70-million-americans

[147] O. Solon, "Facebook faces $1.6bn fine and formal investigation over massive data breach," Oct. 2018, accessed on 2018-11-10. [Online]. Available: https://www.theguardian.com/technology/2018/oct/03/facebook-data-breach-latest-fine-investigation

[148] U. Iqbal, Z. Shafiq, and Z. Qian, "The ad wars: retrospective measurement and analysis of anti-adblock filter lists," in *Proceedings of the 2017 Internet Measurement Conference*. ACM, 2017, pp. 171–183.

[149] M. Graham, "Digital ad revenue in the us surpassed $100 billion for the first time in 2018," 2018. [Online]. Available: www.cnbc.com/2019/05/07/digital-ad-revenue-in-the-us-topped-100-billion-for-the-first-time.html

[150] C. Gayomali, "It would cost each user $232 a year for an ad-free internet, study finds," 2014. [Online]. Available: www.fastcompany.com/3034670/it-would-cost-each-user-232-a-year-for-an-ad-free-internet-study-finds

[151] Amazon, "Alexa top sites." [Online]. Available: https://aws.amazon.com/marketplace/pp/Amazon-Web-Services-Alexa-Top-Sites/B07QK2XWNV

[152] T. Micro, "Site safety center." [Online]. Available: https://global.sitesafety.trendmicro.com