"Senator, We Run Ads":
Understanding Users
Through Targeted Advertisements

Bharath Srivatsan
Adviser: Edward W. Felten

Submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science in Engineering
Department of Computer Science
Princeton University

May 2018

I hereby declare that I am the sole author of this thesis.

This thesis represents my own work in accordance with University regulations.

Bharath Srivatsan

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Bharath Srivatsan

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Bharath Srivatsan

ABSTRACT

Targeted advertisements are central to the functioning of the modern web. Given their significance, many studies have attempted to understand the underlying mechanisms of the online advertising ecosystem. However, most of these works focus on identifying particular targeting behaviors through controlled web simulations. In order to paint a more comprehensive picture of targeted advertising for more realistic browsing patterns, I propose a novel approach that leverages live user data. In this thesis, I investigate what we can learn about advertisers and data flows from observing ads shown to individual users. I conduct what is (to my knowledge) the largest live-user study of targeted advertisements to date, and find that ads can viably be used to reconstruct personal identifiers, sensitive characteristics, demographic details, and user interests, despite the potential confounding variables that typically complicate such measurements.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help of many individuals.

To Professors Felten and Kernighan: you were the legends I'd heard stories about when I was just a freshman choosing my major. It has been an honor and a privilege to work with you. To all of the researchers who offered me help and guidance – Professors Narayanan, Mayer, and Stilz at Princeton, Professor Dali Kaafar at Macquarie, and Jaideep Chandrashekar at Technicolor – thank you.

To the individuals who contributed to this thesis – thank you for investing your time in this project. This is as much my work as it is yours. To my thesis fairies and copyeditors – Sunita, Becca, Crystal, Peter, Usama, Manisha, Wendy, Stephen, and An Lanh – this would not have been possible without your love and your feedback. To Vishan and Simran – your guidance and support helped this COS major pretend to be a statistician. To everyone who installed my extension and participated in my study – thank you for trusting in me.

To the communities on campus that have made Princeton my home (PDP, Tower) – you will be who I remember when I remember this school. To my friends, classmates, and mentors – you shaped my experience at Princeton, and in doing so shaped who I am. To my roommates – Waqa, Soham, Usama, and Manisha – without you, I'd have finished by March. Thank you for not letting me.

To Appa, without whose guidance I would not be here; to Amma, without whose support I would not be here; to Sunita, without whose love I would not be here: thank you.

To Amma, Appa, and Sunita.

CONTENTS

List of Tables

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

On April 10th, 2018, Mark Zuckerberg, the CEO of Facebook, testified before the U.S. Senate. Zuckerberg's appearance on Capitol Hill was the natural next step in a series of scandals that had rocked both the company and Silicon Valley writ large [1]. The most recent of these involved Cambridge Analytica, a private company that had used a viral Facebook quiz to scrape the profiles of over 87 million users [2]. Facebook's vulnerability to Cambridge Analytica arose out of specific flaws in the company's data access policies. The pattern of personal data misuse and exposure still seen throughout the industry, however, is indicative of a more fundamental issue.

One explanation for these systemic concerns points to the world of online advertising. Much of the modern web runs on ad revenue. Facebook and Google, two of the largest companies in the world by market cap, rely on serving ads, as

do many millions of smaller sites that make money by displaying ads. Some argue that this model is built to fail with regard to user privacy. Through the lens of this argument, the Cambridge Analytica scandal was a natural byproduct of Facebook's dance between promising advertisers ever more nuanced pictures of users and promising users ever more advanced privacy protections. At some point, it would stand to reason, the company would go too far in either direction, incurring the wrath of either corporations or consumers.

I do not hope to assess whether advertising really is responsible for such user privacy violations. At the very least, though, it seems clear that online advertising is closely linked to online privacy. Advertising, after all, is both a source of user data and a motivation for many third party data transfers. As digital ad spending continues to skyrocket (see figure 1.1) this revenue model will only grow in importance. Compounding this worry about commercial incentives to overlook user privacy, new techniques have made it easier than ever for advertisers to compile vast collections of data on individual users. Digital ad spending is a lucrative domain rooted in an unprecedented level of insight into individual users, and as competition for ad dollars intensifies, we as a society ought pay close attention.

Moreover, in many ways this ecosystem has grown more opaque over time. New tracking and tailoring mechanisms are continually being developed and deployed, and disclosures on their use are few and far between. Facebook and Google alone captured over 63% of US digital ad spend in 2017 and yet there is much we do not know about how this duopoly aggregates and uses the data it collects [3]. While this continues to be a popular space for research, for reasons I go on to identify most past works simply focus on controlled, measured simulations aimed at teasing out specific features of the online ad ecosystem.

It is in this climate that I present my thesis, a study of targeted advertisements and their implications for user privacy. I focus on a single organizing

research question: what can we learn about advertisers and data flows from ads targeted to individual consumers? Within this broad space, I identify subdomains of questions that, for various methodological reasons, have received comparatively less attention in past analyses.

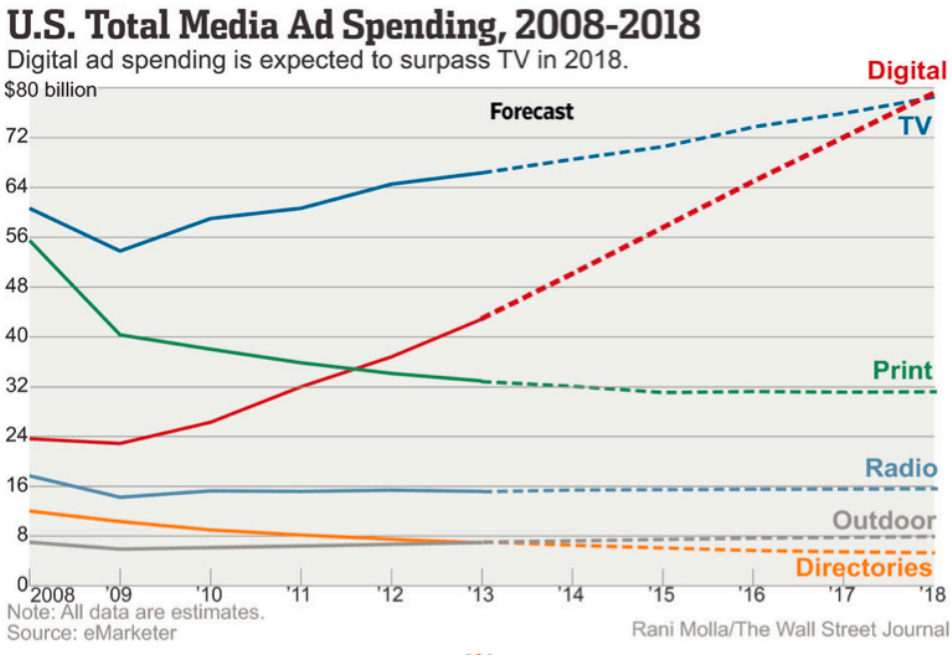


Figure 1.1: *Ad spending by media type over time* [3], [4]
N.B: *Digital ad growth outpaced estimates and overtook TV spend in 2017.*

More explicitly, my thesis contributes to existing literature in a few ways:

- I conduct what is, to my knowledge, the largest live-user study of targeted advertisements to date (and the largest study of its kind for advertisement privacy issues on Facebook);
- I demonstrate the ability to learn deep levels of information about users in the wild - including personal identifiers, sensitive preferences, demographic categories, and interests – solely from observing ads targeted to them;

- I conduct a literature review of over 40 papers in this domain to map out past work and identify potential avenues for future research; and,
- I outline and justify a three-pronged ethical standard with which to evaluate corporate stewardship of online advertising data

I begin in Chapter 2 by providing background information on online advertising and the various techniques that are commonly used by advertisers. In Chapter 3, I conduct a literature review and outline some of the most important papers in this space. Chapter 4 serves as the motivation for my thesis, analyzing in greater depth the reasons for studying this domain and justifying some of the methodological decisions I make (including my choice to study live users). In Chapter 5, I describe my approach, focusing both on how I collected data and how I tackled each of my six specific research questions. Chapter 6 illustrates the results of these methods and includes a discussion of my datasets, my analyses, and various case studies I examined. Finally, I conclude in Chapter 7 with my most salient conclusions and avenues for future research.

CHAPTER 2

ADVERTISING ONLINE

Targeted advertisements occupy a unique space, both having enabled and having been enabled by the growth of the internet. Their foundational intuition is simple enough: unlike old forms of static, mass-distributed advertisements on TV or in print media, the internet allows advertisers to serve distinctive ads to distinct individuals. Over the last two decades, this ecosystem has grown ever more complex. Ad networks and ad exchanges, engaged in an all-out war for customer conversions, have developed increasingly refined techniques to track and understand internet users.

This process was hastened by the development and widespread adoption of Real-time Bidding (RTB), a technology that allowed companies to compete for ad space dynamically. When regular users today access any of the many RTB-enabled websites across the internet, their gazes are instantaneously appraised. Advertisers

of varying sizes and industry verticals silently bid for the right to serve their own content to each new observer; before the website fully loads, a winner is selected and an advertisement is spliced in [5]. RTB is but one of a host of technological developments in the advertising space, but its importance cannot be understated: it enabled advertisers to tailor ads hyper-specifically to the potential customers they wanted to reach. Ad distributors quickly realized that corporate ad budgets would be directed to the services that best leveraged RTB, prompting the distribution of ads that were directed in content, style, and scope as never before [6]. This ability to customize, combined with a newfound capacity to learn ever more information about viewers' preferences, interests, and habits, has contributed to the technique's value (and its meteoric growth) [4]. Advertising providers can now charge companies top dollar for the ability to display messages to prime customers.

In this chapter, I provide a high-level overview of the targeted advertisement space, discussing the key tools in operation. I begin by describing the various types of agents involved and the ways that they've interfaced to build this ecosystem. I then outline some of the methods used to track users as they browse the web, touching on both traditional cookie-based approaches and more novel modifications. Finally, I describe some of the tailoring approaches used by advertisers to personalize their ads in flight.

2.1 Advertising Agents

At the turn of the millennium, global digital ad spending sat at just above $9 billion US dollars. In 2017, that same figure had increased by over 2,240%, marking the first year digital spending overtook its TV counterpart [4]. This breakneck growth has been accompanied by an explosion in companies offering online ad solutions. While the inherent complexity of the space makes any attempt

at simplification difficult, there are two archetypical channels for ad distribution widely in use today.

The first makes use of ad networks. At a basic level, ad networks contract with site publishers to pool inventories of website space that they can then sell to advertisers. These networks act as the intermediaries between sites and brands, attempting to match as many 'impressions', or ad views, as possible [7]. Ad networks can come in a variety of forms – some focus on particular verticals (selling only to specific kinds of businesses), while others cater to different classes of ad purchases, focusing either on premium publishers or on massive-scale, low-quality spots. Still others differentiate themselves on ad formats, such as by specializing in mobile or video ads. Some companies, attempting to capture as much of this space as possible, own multiple ad networks that capitalize on different types of consumers; Google controls AdSense (for general ads), DoubleClick for Publishers (for premium content), AdMob (for mobile ads), and more [7], [8].

Despite the vast array of ad services, targeting ads to optimal users on traditional ad networks can still be difficult. While networks offer the ability to cater ads to sites with differing user profiles, more granular tuning is often impossible [7]. Ad exchanges have recently grown in prominence as a potential solution to this issue (and represent the second major channel for digital advertising). In a programmatic ad placement, advertisers connect to Demand-Side Platforms (DSPs), while site publishers list available ad space on Supply-Side Platforms (SSPs). These DSPs and SSPs then come together in an ad exchange, which hosts an instantaneous auction for each potential impression. Sometimes, by integrating with Data Management Platforms (DMPs), SSPs can aggregate more information on a viewer's characteristics to display at the auction, which can thereby attract more demand-side competition and higher prices. Like ad networks, some ad exchanges limit themselves to only some kinds of advertisers [9].

Seeing the benefits of RTB, some ad networks have built in more dynamic placement features. In the meantime, some SSPs are offering features previously found exclusively in ad exchanges or networks [9].[1] As this convergence continues, drawing clear lines grows increasingly difficult. For the purpose of this thesis, I focus on ad networks and exchanges with robust ad tailoring and targeting functionality built in. I also use the term ad server or ad agent to method-agnostically refer to an ad intermediary [10].[2]

2.2 Tracking Methodologies

When targeting advertisements, data are crucial. More data mean a more accurate picture of potential consumers, a more accurate picture means more specific ads, and more specific ads mean more purchases. One crucial method for gathering such data utilizes browser cookies, small pieces of text that allow websites to flag visitors and retain information about them for future visits. Browser cookies are not inherently advertising-oriented: cookies are also used to personalize sites and improve browsing experiences. A travel website, for example, could drop a cookie onto a user's device that indicates their chosen language and country. The next time this user accesses the website, their browser will automatically send this cookie back, allowing the site to personalize flights and deals to the user's preferences. Publishers can similarly use such techniques to recommend likely articles of interest to return viewers, while many other sites use cookies to track logged-in users or general traffic (via, for example, Google Analytics) [11].

The examples above are first-party cookies, dropped by the creator of a site to help improve that site's experience. Any dynamically-loaded frame on a website,

[1] Relatively novel solutions like Programmatic Direct blend these lines even more emphatically, allowing automated but direct ad-buying for publishers.

[2] Technically speaking, ad servers are a first-layer mechanism used by publishers and advertisers to manage slots/display ads or track campaigns and aggregate information, respectively.

though, can drop cookies, meaning that entities from Facebook and Google (via their ubiquitous 'like' and share buttons) to ad exchanges (via advertisement frames or pixels loaded on pages) can take advantage of so-called 'third-party' cookies [11]. This means that as users jump around the web, ad servers on the sites they visit update cookies that the sites can later read to get ever more comprehensive profiles of users' browsing behaviors. Given that browsing behavior itself is an astonishingly strong proxy for interests and preferences, this data is invaluable for targeting advertisements [12], [13].

This information can be used in many ways. Most canonically, it can be used by companies to display different products to different people. If Ford and Old Navy both took out online advertisements, for example, they would likely have very different target audiences in mind. If the ad networks they contract with saw that a particular user had a predilection for visiting the websites of Chrysler dealerships, they could make a reasonable inference that the Ford ad would likely be more effective.[3] Advertisements can also be taken out for various ideological causes, making use of the fact that user behaviors are predictive of the kinds of movements to which individuals would be most sympathetic [14].

For users, circumventing cookie-based tracking is simple in theory, but in practice can be difficult to accomplish. Users who set their browsers to refuse or periodically clear cookies, for example, might think that their browsing patterns are not being compiled over time. In response, ad intermediaries have developed more intricate cookie constructions that are resilient to such attempts. Ad exchanges don't typically store full records of user browsing information on the cookies they place (due to space constraints). Instead, they retain this information

[3] In practice, RTB would act as an intermediary step that takes the process of making such inferences out of the hands of ad networks. Both companies would be allowed to bid for ad space shown to this customer; Ferrari, presumably, would have a higher budget allocated for users who fit such a profile and so would win the auction.

on their own databases. The cookies they drop, then, often consist only of a unique user id tied to a database entry, so users who wipe these cookies won't have wiped the background information that these ad servers have collected about them. If an ad server is able to reconnect a user's actions 'post-wipe' with their previous user id, the deletion will have been worthless. Some agents attempt to do so via a process called cookie syncing, used to cross-reference cookies placed by different ad servers. By communicating with partner systems (whether they be ad exchanges, demand side platforms, or data management platforms), exchanges can develop a table to link user ids generated by different ad servers. Then, if only one agent finds its cookie wiped, syncing with other systems would allow this server to re-identify the user in question based on another's continued tracking [15].[4] A separate workaround involves placing 'evercookies' on users' systems. These cookies are replicated across many different system locations and refresh each other in the event of deletion, making them notoriously hard to conclusively wipe [16].[5] Still more advanced methods purport to follow users across devices using deterministic and probabilistic matching of unique identifiers associated with their profiles [17].

Fearful of over-relying on one type of method, ad agents have also developed a variety of techniques to uniquely identify visitors *without* cookies. One key process by which publishers do so is called device fingerprinting. Since every device behaves a little differently, a website can make innocuous queries about a user's system that, when combined, create a distinctive picture of an individual browser. For example, different devices have different fonts installed – fingerprinting scripts might thus request a list of the font libraries installed on a machine. Devices also draw images differently, so by instructing browsers to render invisible images, a

[4] Such collaborations can include data purchases that have the added benefit of allowing ad servers to see user browsing information from websites they didn't have trackers on.

[5] Most major trackers view this practice as a serious privacy violation and do not engage in placing such cookies. However, through cookie syncing, so long as one tracker in an ecosystem has placed an evercookie, all other trackers may have their original information re-synced after deletion.

publisher might be able to identify quirks that distinguish an individual computer.[6] The capacity for sites to access this sort of information is important and was likely intended innocuously – font requests help to ensure that the user's browser will be able to render website content, while drawing images is a common occurrence on all kinds of sites. Nevertheless, these requests can serve as individual bits of order, chipping away at the randomness assumed in online browsing [18].

A large portion of the most trafficked parts of the web engage in some flavor of tracking – almost 80% of the top 1 million websites on the internet have a third-party tracker from Google. These sites, on average, have around 20 third-party tracking scripts each [19]. This landscape is ever-changing, however. The rise of ad and tracking script blockers like AdBlockPlus and Ghostery cost US publishers more than $15.8 billion in potential revenue last year [20]. Google Chrome, the most widely-used web browser, recently released its own built-in blocker for especially bad ads and trackers [21]. As this war continues, tracking techniques will have to continue to evolve.

2.3 Tailoring Methodologies

Over time, the tracking measures described above allow ad exchanges to collect data on users visiting sites serving their ads. These servers then allow ad buyers to use this information to increase customer conversion rates by targeting and tailoring their offerings. Such ad customization can be as simple as refining the text of search ads based on algorithmic A/B tests or as complicated as modifying an advertisement's content in flight to call out potential customers by name [22], [23]. Understanding advertisement tailoring, therefore, requires understanding both the kinds of information used in the process and their different applications.

[6] This technique is known as canvas fingerprinting.

Theoretically, the capacity for ad tailoring is unlimited – with the right background information and an accurate user identification mechanism, advertisements could be personally crafted for individual observers. Even based purely on a one-off visit (ie. with no prior knowledge of a user's interests, demographics, or identifying information), an advertiser could use browser-provided information to tailor ads. For example, a user's IP-address could provide city-level geographic information, while factors like device operating system, browser, and language could provide hints about other user characteristics. With more user data, ad providers could begin to piece together more specific features, including preferences, behaviors, and personal details. At the other end of the spectrum, for a user who engages regularly in tracked online browsing that reflects their identity and past actions, advertisers could go much further: as far as, hypothetically, to suggest by (algorithmically) personalized celebrity endorsement that this user purchase specific products that they had previously viewed [23]. Table 2.1 outlines some of these possibilities. In practice, the personal information available to an ad publisher is broadly constrained by a user's own browsing habits and the information that data providers choose to furnish. Data providers, in turn, often process the raw data they receive from supply-side platforms into thousands of segments such as relationship status, interests, ethnicity, home value, income, connected devices, and more [24].

Once such information about a user is collected, ad publishers can use it in two main ways. First, they can use user data for audience selection, deciding which customers are most valuable (and therefore most deserving of ad impressions). For products that are highly age-specific (advanced gaming laptops, for instance), an advertiser may decide not to show their content to users outside of their target demographic. Advertisers are also increasingly using retargeting, a technique used to display ads for products that users previously viewed. In this way, an online

shopper who 'carts' a pair of shoes but does not buy them could be reminded of their potential purchase with ads for the same shoes on the sites they jump to next [25].

Second, advertisers can use such details to personally customize their ads. Ad personalization is an increasingly prevalent strategy used by brands to both optimize content and break through the clutter of irrelevant messages on the web. Google Adwords, for example, allows advertisers to include 'ad customizers' in their text ads. These take the form of placeholders in ads dynamically filled in based on a comprehensive mapping of various user attributes to variable values [26]. This way, an ad for a sale could highlight the discount associated for the specific products a user is most likely to want. There are limits to such customization, though; in order to protect users and reduce intrusiveness, some firms (including Google) do not allow ad publishers to directly include personally identifying information. A hyper-personalization strategy aimed at singling out a user by name, email, or id number would likely be rejected by the search giant [27]. However, across the ecosystem, conditions are far more murky. Different publishers accept different degrees of customization, and so ad personalization techniques are likely to continue to deepen in scope and focus.[7]

[7] Interestingly, one of the most serious critiques floated against ad personalization isn't ethical in nature but economic: some recent studies appear to show that hyper customized ads actually decrease user purchasing intentions because they prompt feelings of intrusiveness [28].

Table 2.1: *Categories of information used in tailoring advertisements (c.f. [23])*

Session Type	Information Available	Example Usage
One-off	• Location (via IP address) • Device details • Host website information • External conditions (time, weather, world events, etc.)	*Tailoring an ad for a service by announcing operations in <user city>*
Tracked browsing history	• Needs/wants (via retargeting) • Demographic information (age, gender, income, etc.) • Interests (sports, tech, etc.)	*Resurfacing an ad for a product a user placed in their cart (but didn't buy)*
Tracked preferences	• Advanced demographics (political leaning, brand interests, etc.) • Niche interests (specific artists, hobbies, sports teams, etc.)	*Targeting ads for a political party to users interested in similar politicians*
Tracked behavioral patterns	• Behavioral history (purchase record, location history) • Purchasing intent (via keyword search history) • Connection history (preferences or patterns of friends and influencers)	*Recommending a particular brand to users searching for a product category who have previously purchased from a competitor*
Tracked personal details	• Personally identifying information (name, id, email, address, etc.)	*Named callouts: <name>, buy <product>!*

CHAPTER 3

LITERATURE REVIEW

Understanding how practices like tracking and tailoring work in theory is one thing; understanding how and when they are applied in the real world is another altogether. In this sense, ad exchanges may seem like black boxes to the outside world. Indeed, as advertisers have grown increasingly sophisticated, the methods they use have grown increasingly opaque; public disclosures from these entities about the ways they operate are rare [29]. Fortunately, though, online advertising is a well-studied space. Researchers have developed a host of useful techniques to survey the web in order to identify patterns in and derive conclusions about ad interactions. In this chapter, I outline relevant work done in the domains of web privacy and online advertising tracking. I first discuss past studies on online

tracking and advertisements and then investigate research into defense mechanisms for consumers.

Before I examine these past works, however, it is important to note what works I will *not* discuss. For my review of privacy studies, I focus exclusively on passive research about online display advertising. Many studies, though important and relevant to the domain writ large, do not fit this criteria. For example, [30], [31] investigate the possibility of using microtargeted ads to identify users, while [32] demonstrates a method of polluting personalization mechanisms to upset targeting (including of ads). These papers focus on *active* threats, ie. potential attacks that directly engage with or disrupt advertisers' internal mechanisms. However, since passive (observational) studies are more relevant for painting a descriptive picture of online advertising and can illuminate threat vectors that are particularly hard to detect, I focus on passive methods in this thesis and exclude active approaches from this review. I exempt this condition for my analysis of defense mechanisms, since active defense techniques can be just as useful to consumers as passive ones. Other studies, including [33]–[35], focus on privacy violations due to web searches, email scanning, and collaborative filtering (respectively). While such works help paint an overarching picture of privacy in the modern web, they are less relevant for understanding the privacy landscape of targeted display advertising specifically. For the same reason, I ignore specialized investigations of privacy violations on social networks like [36]. Finally, I exclude studies like [37] that focus exclusively on mobile web tracking.

3.1 The Advertising Ecosystem

Even given the above exclusion criteria, there are plenty of papers that study the online ad ecosystem. I organize them based on the targets of their analyses, subdividing the space into studies of user tracking data, data outflows, data usage,

and data inflows. Broadly, reports on tracking data aim to understand how user behaviors are traced through different sites across the web. Data outflow research focuses on investigating how and what data first and third party advertisers send to each other. The data usage category includes research into how various user characteristics are used in constructing and targeting ads. Finally, data inflow studies analyze ads to see what kinds of data are present in them. These categories are organized in a chronological sense; users are first tracked around the web, browsing sites and building profiles that then flow to third parties and ad servers. Next, these ad agents use particular user characteristics to construct ads that ultimately flow back to users. In using this breakdown, I build upon a framework established by Englehardt et al., which defines web privacy research as attempting to measure or infer data collection, data flows, or data usage [38].[8] Table 3.1 provides an overview of these targets.

3.1.1 Tracking Data

Tracking-related research comprises a large part of the web privacy measurement space. Roesner et al., in 2012, developed a method for detecting five specific kinds of third-party trackers and then applied it to simulations of users visiting sites [39]. In doing so, they were able to reach various conclusions about trackers in the wild. For example, they found that there are often many trackers on a single website and that the top few tracking companies place the vast majority of tracking cookies overall. Though this study was limited in scope (both in terms of the number of tracking methods analyzed and in terms of the site survey size),

[8] Beyond nuancing [38]'s exploration of data flows and adding new areas of analysis in my discussion, I also limit the scope of their framework to only those questions that have applications to online advertising. Finally, I investigate additional works that fit in this picture, including papers released after [38].

Table 3.1: *Overview of ad ecosystem research subdomains*

Target	Overview	Sample Questions	Select Papers Discussed
Tracking	How are users followed around the web?	How prevalent are trackers/cookies? Who places these trackers, and on what sites? Which cookies are used to identify users?	Roesner et al. Englehardt and Narayanan Macbeth Acar et al., 2014
Outflows	What data is sent to third-party services?	What personal data is leaked by first-parties? What information is stored in outgoing cookies? What data flow techniques are commonly used?	Krishnamurthy et al. Meyer Olejnik et al. Acar et al.
Usage	How is user data used for ad targeting?	How does location affect ad targeting? How do user demographics affect ad targeting? How do user interests affect ad targeting? How does browsing history affect ad targeting? How do user attributes affect ad content?	Datta et al. Castelluccia et al. Barford et al. Wills and Tatar Liu et al.
Inflows	What data is latent in incoming ads?	How often are ads about sensitive content? Do ads commonly contain personal identifiers? Do ads often link to malware or through HTTP?	Datta et al. Castelluccia et al.

[19] replicated these results (and others) on a census of the top million websites on the internet. More recent studies have pushed this sort of research further - [40] extracts third-party trackers embedded in one billion websites saved in the Common Crawl dataset. While this work is able to more conclusively establish the long-tail distribution of trackers and answer interesting questions about tracking across countries and on privacy-sensitive websites, its reliance on the static Common Crawl dataset means that it likely missed out on some tracking methods dynamically engaged on live site visits. [41] fills this hole by studying 850,000 users of Ghostery's opt-in GhostRank feature. By examining over 440 million real user page loads, this study is able to investigate more natural user flows and interactions than its synthetic predecessors. While each of these studies contribute slightly different conclusions, they agree on a unified picture of cookie-based web tracking as incredibly prevalent.

Still other studies focus on more complex forms of web tracking (see §2.2). [42] analyzes the code of various device fingerprinting libraries to establish both that the methods therein would be effective on popular modern browsers and that they are in fact already being used by some sites on the internet. Parallel to the aforementioned developments in cookie-tracking analyses, measuring device fingerprinting soon become algorithmic and applied to huge crawls of sites. [18], [43] crawl 1 million/100,000 sites (respectively) to survey the prevalence of font-based fingerprinting, canvas fingerprinting, and/or evercookie placement (see §2.2). While these works suggest these more complex scripts are less common than traditional cookie-based approaches, the inadequacy of current consumer defenses against such tactics leaves room for concern. Finally, research in this subdomain can also focus on understanding defensive measures employed by advertisers and ad servers. For example, [44] surveys the top 100,000 websites to identify anti-adblockers that modify site content if in the presence of an adblocker.

3.1.2 Data Outflows

Understanding outflows involves understanding how, what, and where information is sent. This task is typically applied to one of two avenues: outflows from browsers/first-party sites to third-parties, and flows between servers (eg. cookie syncing/matching[9]). Monitoring outflows from user browsers isn't a complex undertaking (since researchers can just track outgoing HTTP requests), and simply watching these communications can lead to some surprising findings. [45], [46] observe (among other things), that sensitive or private strings sent to healthcare and flight booking sites were leaked to third parties by nine of the top ten websites in their respective categories. [46] finds transmissions of names, email addresses, and/or phone numbers to third parties just for viewing ads or changing basic settings. As the internet has moved towards an all-HTTPS ecosystem, such measurements may grow far tougher; interpreting cookies and web requests may no longer be as simple as sniffing for plaintext information. Unfortunately, though, this does not imply that these sharing behaviors will cease. Networks that agree on encryption in advance can use the same sharing methodologies, just with encoded information. Just as worryingly, even encrypted cookies can be used to track users: [47] outlines a method of surveilling users by cross-referencing cookie placements.

Cookie syncing is theoretically trickier to study because it involves flows of information between two third-party actors. However, many cookie matching protocols operate via the user's browser as an intermediary – that is, ad frames often include scripts that instruct *browsers* to send cookies and exchange ids to partner services. In 2013, [48] used this fact to perform an astonishingly in-depth analysis of cookie syncing. That work found that cookie matching happens frequently and for significant proportions of users (and went as far as to estimate

[9] I use the terms cookie matching and syncing interchangeably here; they refer to the practice outlined in §2.2.

the prices companies were paying for user profiles!). Acar et al. exploit the same vulnerability to investigate userids synced with cookies [18]. It is important to note here that there's no easy way of examining private data transactions or sales between third-parties that don't use end-user browsers as an intermediary. Data aggregators often sell batches of user histories to ad networks or the like; studying these flows is incredibly challenging.

3.1.3 Data Usage

Given the diverse array of data that is available to advertisers, it should come as no surprise that studying how these data end up being used is a large and open-ended field. Indeed, this subdomain has grown increasingly popular as an area of study in recent years. In light of recent developments around consumer demands for transparency in data usage by big tech firms, understanding how information is used in ad construction will likely become even more pressing [49].

As outlined in Chapter 2, user information is used to maximize conversions and revenue by either tailoring content or improving ad targeting. With respect to the former method, [50] found that user browsing behavior can be used to price discriminate on the web, a finding empirically borne out by companies like Orbitz charging Mac users more for hotels [51]. Perhaps more pernicious is Sweeney's finding in [52] that racially-associated names can produce Google ads tailored to negative stereotypes about those races. Content tweaks are also popular; ads, often for unsavory or sensitive services, have long included user details like home cities in callouts. As briefly discussed in §2.3, content customization in ads is theoretically limitless; even the campaign A/B tests performed across the web are a form of tailoring. This open-endedness, however, makes this domain particularly difficult to study on an expansive scale. It is hard to establish causality in ad customization across general censuses of ads due to the presence of various confounding factors

(including the presence of A/B testing). This is especially true for live-user studies, in which variables may differ so drastically that large datasets of ads from the same providers (needed to detect causal customization) are all but impossible to collect. As a result, most of the above papers illustrate the existence of tailoring anecdotally, on restricted domains, or for specific use cases.

User profiles are also used to target ads. Since bidding on ad exchanges often involves complex, real-time, algorithmic pricing decisions, researchers cannot easily access the key determinants that contribute to any particular impression. With that said, studies have been able to simulate user profiles to observe ads and make inferences about data usage. These works can be broken down in three ways – by the information they attempt to test for, the methodologies they use, and the ad providers they investigate.

Studies in this domain attempt to test for a variety of informational inputs in ad generation. Location, for example, can be used effectively to target ads [53]; studies could try to identify location-based targeting by simulating identical user profiles hitting websites from different IP addresses. The bulk of research here, though, is devoted to understanding how demographics and interests affect ad targeting. [54] varied user characteristics like age and gender and searched for statistically significant differences in ensuing ad impressions. Datta et al. found many sites that targeted ads to women, and (separately) that ads for job opportunities shown to women offered lower salaries on average than those shown to men. Other studies simulate different user interest profiles, generating canonical users interested in 'Arts' or 'Shopping' or 'Finance' and observing what types of ads are targeted to them [55]–[57]. [56] used this method to generate a profile heatmap illustrating how different interest profiles attracted ads from interest categories other than their own. Meanwhile, [54] observed ad targeting based on simulated interests that didn't appear on users' Google Ad Preferences Manager,

suggesting that interest-based ad targeting might be more nuanced than the categories that Google assigns.

Generally, researchers in this space employ a three-step approach. First, they generate web browsing profiles either by setting values explicitly via an ad preferences site ([54], [57]), by browsing sites related to particular interests or categories ([54]–[56]), or by re-creating the browsing behavior of real users ([55], [58]). Having built different profiles with different characteristics, researchers then typically have crawlers visit many websites and collect the ads they find. Then, by noting the differences in what ads are collected for each simulated profile, researchers can make inferences about causal effects [59]. Some works vary pieces of this archetypal architecture: [60] switches stage one by generating profiles from user behaviors on Gmail, Amazon, and Youtube.

Most papers in this space concentrate on ads distributed by Google, since the Google Ad Preference Manager allows users to set and observe their own interest and demographic designations, and since Google ads are so widespread. Comparatively less research has been done outside the Google ecosystem. [60], as mentioned, examines Amazon, while [57] performs a limited Facebook study. Ultimately, the difficulty of creating and populating realistic mock user profiles on Facebook has largely slowed research on the platform.

Of note here is the striking lack of live user experiments. I use the term 'live user study' to refer to studies that collect data directly from users fully in the wild. Though there are many works that capture real users' browsing histories to feed into simulated modules, the 'live user data' they collect only serve as an intermediate stage in their analyses. Live studies under my definition stand to benefit from the factors I will outline in §4.2; simulated studies (even with live components) do not.

Assessed as a whole, these papers paint a compelling picture for future work in this space. On the whole, ad networks seem to strongly discriminate on what ads they target to users based on a range of user characteristics and interests, including some not reflected on Ad Preference Managers. These behaviors occur in ways that are not completely understood, in a domain that is continually evolving, and that has been researched largely through simulated user orchestration.

3.1.4 Data Inflows

Comparatively little research has been done in what I term the 'data inflows' subdomain. This may be because identifying personal identifiers leaked through sites or classifying sensitive ad topics is an open-ended task most compelling in a real-world setting, rather than on simulated profiles built to draw out such results. That isn't to say that identifying such violations is impossible – to the contrary, [54] finds that after accessing sites about substance abuse, disabilities, dating interests, or weight loss (all potentially sensitive topics), users received statistically significant increases in ads from those categories. However, this study was severely constrained – researchers manually picked out a few sensitive categories and investigated only those effects. Meanwhile, [55] notes from their wide-ranging study that the "dating" category appears in ad profile reconstruction, but does not attempt to extend this analysis more broadly to reach conclusions about other kinds of sensitive ads.

3.2 User Defenses

Researchers in the online advertising domain have also described and evaluated various user defenses against these tracking and sharing behaviors. Some of these works describe new models for the space as a whole - [61], [62] describe

Adnostic and Privad (respectively), new online advertising systems that will be more respectful of user privacy. Still others investigate novel techniques that could potentially be implemented by companies to improve user browsing experiences. [63] presents Perceptual Ad Blocking, a framework to identify ads through contextual characteristics that will allow users to block ads and ad tracking more effectively. FPRandom, from [64], attempts to disrupt device fingerprinting by randomizing different browser aspects.

A third class of papers in the user defense space focuses on practical measures at an individual level. These works evaluate the success of various off-the-shelf options in order to make prescriptions on how users can maximize their privacy on the web. One of the older papers in this space, [65], analyzed the effectiveness of browser-based instructions on dissuading tracking. The authors found that while opt-out cookies and blocking methods limited behavioral advertising, Do Not Track headers weren't effective. More recent works investigate these tools and others on larger datasets; [66]–[68] assessed the effectiveness of various adblockers. [68] was largest by number of sites hit (100,000 in total), while [67] tested a wider range of extensions and browsers. The latter study assessed both tracking reduction *and* page quality preservation in evaluating extensions against one another. This type of dual mechanism is important as it recognizes the inherent tradeoffs that users realistically make when deciding on which tools to engage. Unfortunately, these studies do not converge on a single, effective recommendation for users. Figure 3.1 provides a general overview of tracking mechanisms and corresponding user defenses currently available on the web.

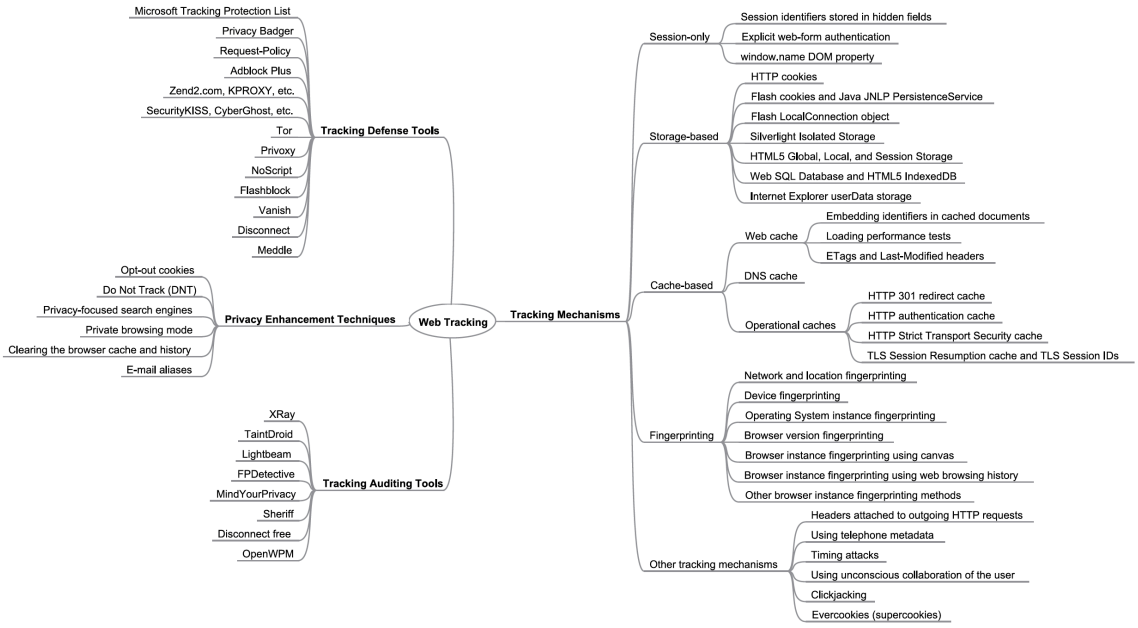


Figure 3.1: *Overview of tracking mechanisms and user defenses* [69]
N.B.: *This chart is not comprehensive – it does not, for example, include popular tools like Ghostery*

CHAPTER 4

POSSIBILITIES, THREATS, AND IMPLICATIONS

Chapter 3 surveyed past work done in the online advertising space. As I explain below, however, those works aren't fully comprehensive. There are still significant holes in our understanding of both the targeted ad ecosystem broadly and Google/Facebook specifically. Of particular interest is 'leaking data', a blanket concept that refers to the information latent in ads that can reveal user characteristics. This leakage can be explicit (an advertisement could mention a user's name in a callout) or implicit (the choice of ad targeted could 'leak' user interests if put through more advanced analyses).

My thesis hopes to add slightly more color to our understanding of this complex ecosystem by focusing on leaking ad data. In this chapter, I outline and justify my overarching research goals, beginning by motivating an examination of

leaked data. I then provide an overview of the specific questions I plan to research, and conclude by establishing a three-pronged ethical standard with which to judge large data purveyors.

4.1 Motivation

Investigating leaking data is important for a variety of reasons. In this section, I identify three – gaining a better understanding of advertisements, learning about Google and Facebook, and evaluating threat vectors associated with these data.

4.1.1 Examining Advertisements

As hinted at in sections §3.1.3, §3.1.4, and §3.2, there are still several interesting open questions in the online ad privacy space on which researchers have not yet reached a consensus. Many involve, at least on some level, interpreting the data latent in ads. We may wish to measure, for example, the rate of explicit ad preference leakage for a wider range of sensitive topics and issues than has previously been examined. Similarly, we might want to assess how well techniques for demographic profile categorization or user interest reconstruction work on users in the real world. Given new developments in the war between trackers and adblockers, we may also want to revisit prior findings on adblock outlooks and effectiveness.

These sorts of questions are vital for painting an accurate, up-to-date picture of user privacy considerations in online advertising. We may want such a picture simply because it could give us insight into how an immensely important system operates. Alternatively, given the inherent complexity of the space, we might want to develop mechanisms by which users could reverse engineer data usage. Such a capacity would allow consumers to gain more granular information on how their

data is deployed, allowing them to make better decisions on what privacy measures to engage and what services to continue using.

4.1.2 Understanding Market Leaders

Large data collectors like Google and Facebook have access to huge swathes of information about their users. Beyond characteristics that consumers self-report upon signup, usage patterns can be an effective proxy for user preferences and attributes. Facebook likes alone are a strong predictor of gender, race, sexual orientation, and even religious background; a University of Cambridge study demonstrated that a model given only a record of past 'likes' could predict participants' personalities more accurately than their friends could [12], [13].

Two factors in particular make this degree of access concerning. First, these services have almost monopolistic control over their respective industries. In the US, 89% of internet searches run through Google, while 95% of young adults use a Facebook product [70]. As noted in Chapter 1, these two firms together account for 63% of all online ad spending [3]. These statistics illustrate the market power concentrated in these firms as gateways to the modern internet. Indeed, given the pressing network effects that these companies enjoy, even users who may not otherwise be willing to participate in wide ranging data collection may be forced to do so to engage in the online economy [70]. Even when opt-out mechanisms are available, general users lacking technical skills may be unable to engage them. Second, a lot of this data collection occurs in the background; beyond a Terms of Service contract customers must sign when setting up their accounts, these companies don't typically publicize the scope or regularity of their data collection, not to mention the fact that even users who don't have accounts might be being tracked [2]. More worryingly, even if users are broadly aware of, for example, Facebook's use of 'likes' in targeting advertisements, they may not realize the full

revelatory power of these data. Users may also not fully understand the backend agreements between these companies and those taking out advertisements on their networks [1].

Taken together, these factors mean that studying Google and Facebook is a particularly important subdomain for new research. There are a variety of questions we might like to understand about these market leaders – to begin, how do different sources of user information get used in ad construction, how accurate and comprehensive is this information, and how much do users trust these companies with these data? To answer them, we would need to investigate data leaked from these ad agents in conjunction with users' outlooks and beliefs.

4.1.3 Evaluating Threat Vectors

Thus far, I have focused on why understanding the behavior of authorized ad agents is an important task. Just as importantly, however, if not more so, is assessing how malicious third parties could interface with this system and potentially harm others. I identify two potential threat vectors in the context of leaking ad data that are important to guard against.

The first involves ad eavesdropping. Ad networks lagged behind the HTTPS revolution. For years, websites were loath to enable SSL encryption because they risked ad revenue drops of as much as 30-75% from unencrypted ad demand that could no longer be served [71], [72]. As a critical mass of publishers, encouraged by browsers like Chrome and Firefox, made the switch, the risks of user traffic snooping dropped dramatically. What of ads, though? Advertisements that distribute content via or link to sites through unsecured protocols may still be exposing users to serious privacy risks. This may occur on unencrypted sites, but

might also happen via ad links on SSL-protected pages.[10] Imagine, for instance, if personal identifiers, sensitive characteristics, demographic details, or user interests could be reconstructed from ads, and that these ads sent click cookies over HTTP. This would mean eavesdroppers on public WiFi networks, routers, or other communication channels could be listening to worryingly comprehensive pictures of users. Understanding leaking ad data, then, is critical for examining the risks associated with malicious eavesdropping.

The second threat vector deals with partially-obscured web observation. Many institution-provided devices (eg. company phones or laptops) compile full logs of user behavior. While most of these operate in all contexts, some activate only in corporate settings [73]. For users in the latter camp, private characteristics or interests might be revealed through leaked data from ads seen while at work or school. Though we may assume that the majority of consumers avoid this risk simply by using personal devices for personal use, this isn't necessarily the case. Google Chrome allows users to log into their accounts across different devices, and encourages them to do so by offering improved functionality. Once logged in, consumers can sync bookmarks or passwords and easily sign on to other Google services. Unless an individual opts out, though, their ad profile is also shared between these devices, potentially leaking information to corporate observers. This sort of partially-obscured observation can be incredibly problematic. Imagine, for instance, an individual whose sexual orientation or religious background is carefully concealed while at work or school. If monitors began observing ads that strongly

[10] Ads served via HTTP on HTTPS sites create mixed content warnings on many browsers, meaning publishers often require ad content to be encrypted. However, these ads might link to content that isn't protected. Furthermore, even if the companies taking out ads have encrypted sites, many ad networks might redirect to them via link-click trackers that aren't.

suggested those characteristics, users could face censure or, in some contexts, far worse.[11]

4.2 Research Overview

To tackle some of the above issues, I conduct a live user study of Google and Facebook-owned ad exchanges. I limit my study to Google and Facebook for two reasons. First, as described in §4.1.2, the sheer market power of these firms makes them particularly interesting targets of research. Due to their dominance, they interact with users in a variety of unique ways (not least the fact that their ecosystem of products gives them access to incredibly private and nuanced data); focusing on the two will allow me to investigate the implications of this in more depth. More practically, Google and Facebook are comparatively easy targets for a live user study – they each have huge user pools and voluntarily disclose information like interest and category classifications through ad preference managers. Further, given their wide reach, I do not worry that constraining my study will affect my ability to collect sufficient data.

I propose a live user study for a variety of reasons. Perhaps most saliently, I hope to reach realistic conclusions about advertising behavior in the wild. The threat vectors I outlined above ought to be real worries for real users, and while a simulated assessment might be able to validate their possibility, demonstrating these risks on live (anonymized) users will lend a more significant weight to this discussion. More broadly, user behaviors vary sharply – from browser choice down to cookie use frequency and specific browsing habits. In many ways, simulated

[11] [55] notes an additional, but related, worry to motivate their work. They paint a picture of ad observation by government agents like immigration authorities, who may not have access to browsing histories. Should they gain access to user devices (c.f. 'digital strip searches' [74]), they may observe private behaviors far out of their jurisdictions. We needn't go this far - even information leakage via ads to individuals standing over users' shoulders as they browse should be worrying.

studies cannot capture that diversity. For one, while browser orchestration via modules like Selenium can simulate page accesses more realistically than headless web requests, the complex programs that ad networks run might be able to flag such attempts. There is no guarantee that fingerprinting scripts cannot recognize that different browsers simulated from the same server represent the same 'user' and respond by merging ad profiles and sullying results.

A live user study also opens up a variety of interesting research avenues that were previously out of reach. Most importantly, live user studies can analyze Facebook ads in a far more rigorous sense. Simulating Facebook profiles and behavior realistically is incredibly difficult given the importance of friends in diffusing content and ads. Attempts at orchestrated Facebook analyses, therefore, run the risk of being unrepresentative and ungeneralizable to real users. A live user study can also survey users on their privacy outlooks and practices, opening up new levels of analysis. Not only will my research examine the technical circumstances of leaking ad data, it will also suggest possible implications based on participants' subjective feelings on the matter. Finally, studying live users opens the door to unintended research conclusions that can reveal interesting behaviors that weren't originally being investigated.

The main drawback of this choice is that because of the natural complexities of live user studies I will not be able to conclusively "answer" any individual question for users of Google and Facebook generally.[12] With that said, I perform my study with the hope of reaching reasonable conclusions about the following six questions:

[12] For the rest of this thesis, I may refer to 'answering' or attempting to 'answer' these questions. Suffice it to say, I use this terminology with the understanding that my 'answers' are inferences that apply only to the sample sets I survey.

1: **Personal Identifiers**	Is Personally Identifying Information (PII) leaked via targeted advertisements?
2: **Sensitive Sites**	How often do advertisements link to sensitive, malicious, or insecure sites?
3: **Demographic Info**	Can demographic information be reconstructed from targeted advertisements?
4: **User Interests**	Can user interest profiles be reconstructed from targeted advertisements?
5: **Privacy Outlooks**	How do privacy outlooks, privacy practices, and assessed profiles differ across users, and how do these factors affect each other?
6: **Site Differences**	How do Google and Facebook interest profiles differ?

4.3 An Ethical Framework

When we judge companies like Facebook and Google on data protection, we ought do more than measure them against their own standards. The fact that billions of consumers provide these firms with immense amounts of data confers upon them many unique obligations – as Zuckerberg himself concedes [2]. For the same reasons we place additional regulations on public utilities or systemically important financial institutions, or expect that airplane pilots adopt unique duties of care in conducting their work, we must hold these tech companies to higher ethical standards than most. Furthermore, these sorts of ethical responsibilities should apply both to cases of corporate action and to cases of inaction. We would, for example, find it unconscionable if our local bank did not take reasonable precautions to prevent theft, even if it isn't bank employees who ultimately make

away with our money. Given the immense potential for misuse inherent in vast collections of user data, we ought expect the same from online advertisers.

In light of these considerations, the duopoly of the online advertising space ought be measured against three separate standards when assessed on their data use policies. First, we should evaluate the direct legal implications of their practices. If vulnerabilities might mean companies are no longer fulfilling basic legal requirements, they are failing a fundamental ethical duty to follow the law. Second, we ought hold companies to broader industry and scientific norms on data protection policies. This artificially creates competition across market verticals for the adoption of secure standards and reduces our acceptance of runaway data abuse by even monopolistic entities. Finally, companies ought respect reasonable customer expectations of privacy. A 'reasonable expectation' benchmark, loosely defined as the set of protection standards that an average consumer anticipates will be applied to their data, will help bridge the information asymmetries that have arisen in this domain. The latter two criteria serve as proxies for interpreting the "particular ethical responsibilities" mentioned previously. By leaning on the judgements of both industry professionals and regular users, we can create dynamic standards for protection, resilient to future technical developments.

Holding companies to the law is uncontroversial; holding them to the higher standards of industry norms or reasonable customer expectations is less so. These benchmarks, though, have both precedent and ethical relevance. While legal, deviating sharply from norms on data security is doubly abusive – not only could it prompt retaliatory degradations in practices at other firms, it also takes advantage of consumer expectations rooted in broader industry practices. It seems reasonable, therefore, to construct a normative standard that holds companies to each other and finds them ethically delinquent when they reduce the transparency and security of the community as a whole. This isn't a perfect standard; industry

norms provide less guidance for companies that are unique and have no direct peers. However, positive burdens taken on by some trailblazers – advertiser data sharing managers, Do Not Track compatibility, data use and breach disclosures, etc. – can be demanded from others, even those tackling different verticals.

As for reasonable expectations, common law doctrine on privacy provides explicit precedent for its use, often exploring customer expectations as the standard upon which to adjudicate privacy disputes [75]. Courts have repeatedly found individuals and companies responsible for mental suffering and emotional distress caused by unreasonable invasions of privacy [76]. Indeed, reasonable expectations ought to trump strict interpretations of contract law in some cases. Under most ethical systems, we require contracting parties to fully consent to their actions, but a precondition for consent is an accurate understanding of the relevant facts. Given both the threat vectors described above and the legal/technical jargon that often infuses privacy contracts, expecting users to understand the full implications of their consent might be unreasonable. More broadly, given the market power that companies like Google and Facebook have, we ought not let them silently ignore the reasonable expectations of their users without reforming their practices or announcing their methods.

In the following subsections, I briefly outline what each of the three ethical standards might look like if applied to Google and Facebook and where my research questions fit in.

4.3.1 Legal Considerations

The most direct legal obligations placed upon Google and Facebook are those outlined in their own data use policies. Facebook explicitly notes that only "non-personally identifiable information" will be shared with "advertising, measurement, or analytics partners unless you give us permission." [77] Broad

demographic information, therefore, may be provided, so long as it is aggregated in a way that makes it non-identifying. Information may also be transferred to third-party vendors that satisfy Facebook's confidentiality requirements [77]. Google similarly notes that they "may share non-personally identifying information publicly and with [their] partners." [78] Google maintains a set of stringent policies for ad vendors that aim to foreclose some of the risks I go on to explore. For instance, their policies forbid collecting critical personal information over non-SSL protected pages or sharing personally identifying information directly in advertisements or through Google [27].

Some forms of leaked ad data could break the terms outlined by both of these services. Should ads surfaced by these companies directly expose personally identifying information like names or contact information (despite the efforts of Google's vendor integrity checks), they would violate these data use policies. The question of user interests is trickier, as user preferences consist only of broad attributes that may individually apply to large groups of people. However, these characteristics may, in concert, be sufficient to pinpoint specific individuals. Thus, should it be possible to recreate user profiles from advertisements, those ads may implicitly have contained personally identifying information. Whether or not such disclosures will open these companies up to legal action is unclear; at the very least they are in a grey area of legality.

What of broader data protection laws? In Europe, the General Data Protection Regulation (GDPR) will be enforced starting in May 2018, transforming data protection standards in the process. Among other things, this sweeping new piece of legislation enforces stricter consent rules and provides EU citizens a 'Right to Access' their data [79]. The ability to reverse engineer ad targeting mechanisms will be useful in assessing whether companies are following these restrictions. In the US, while the Federal Trade Commission (FTC) has broad authority to punish

companies that do not protect consumer data, enforcement over data breaches or leaks is rare [80], [81]. This, though, only makes profile reconstruction techniques even more vital as a way of galvanizing customers when abusive data usage is suspected.

If the threat vectors I identified in §4.1.3 are exploited, these companies would likely be liable (at least in part) along both these lines. Leaked ad data being used by third parties would violate both companies' terms, and would likely breach GDPR regulations on third-party data usage consent as well. If these exploitations cause real-world harms to users, district courts in the US may also build on precedent in *In re Facebook Internet Tracking Litigation* or *In re Google Cookie Placement Consumer Privacy Litigation* in finding that users' data had some identifiable value that was damaged [82].

4.3.2 Industry Norms

The FTC outlines norms for corporate self-regulation in the online behavioral advertising space. These follow four broad principles: transparency, reasonable security and limited retention, responsible modifications to policies, and affirmative express consent [83]. Private institutions and corporate groups have advocated for similar best practices. The American Advertising Federation's Institute for Advertising Ethics, for example, requires that advertisers respect user requests and never compromise privacy [84]. In concert, these bodies point to the same set of organizing ideas: customers must have information about and control over their personal data, and corporations must be open, responsible, and limited in handling it.

As indicated above, understanding and evaluating profile reconstruction methods is incredibly relevant for assessing transparency and retention of data [54]. Furthermore, by asking users comparative questions about their outlooks towards

these companies, I may be able to draw conclusions about whether users are truly expressing full consent. Finally, the possibility of third parties reconstructing personally identifying information from leaked ad data would mean these companies are running roughshod over these guidelines. Not only would such a vulnerability violate the FTC's first and fourth guidelines regarding adequate and accurate communication with users, it would also prevent firms from doing the requisite work needed to ensure that reasonable security procedures were followed and data was retained only for "legitimate business purposes or law enforcement needs." [83]

4.3.3 Consumer Expectations

In a 2015 study, only 40% of respondents surveyed were even aware that ad providers commonly tracked online behavior, and just over 50% knew that their personal information was regularly being collected [85]. This stood in stark contrast to the proportion of users who approved of or desired such practices – in a separate set of interviews, respondents decried technologies as wide ranging as Gmail's email scanning and cookie-based web history monitoring. The vast majority of them felt that online behavioral advertisements posed a significant privacy risk that they were actively uncomfortable with [86]. Right off the bat, this wholesale rejection of common online ad targeting practices suggests that much more work needs to be done for these companies to fulfill their ethical obligations.

How likely is it that users would overlook the relevant privacy risks of leaked ad data in order to continue being shown targeted, relevant advertisements? Based on the aforementioned survey, not very. Only 23% of users liked receiving targeted advertisements based on their online activities in the first place, while 37% actively disliked them. In fact, over 80% of survey respondents had engaged in some attempts to preserve their online privacy by refusing to disclose certain kinds of

information, deleting cookies, or activating the 'do not track' option in browsers that supported it [85]. A different survey found that 66% of Americans did not want *any kind* of interest-based ad targeting whatsoever [87]. While it is likely the case that users value maintaining access to the suite of free online services funded by ad revenues, it seems as if a lack of accurate information is coloring user consent for these products.

Also worth noting is the trust that Americans uniquely place in the tech sector. According to a Wired report, Google has a 'net favorability' of 82% (88% of respondents viewed the company favorably; only 6% did not) [88]. Even Facebook, the least trusted of the big tech companies (according to a similar survey by The Verge), had a net favorability of above 60% [89]. Compared to most companies in the US, this is abnormally high [88]. Google, Youtube (a Google subsidiary), and Facebook still make up the three healthiest corporate brands in the world, even as they struggle with fake news epidemics and diversity scandals [90]. These carefully cultivated reputations likely affect consumers' expectations of privacy. Users accustomed to the image of technical competence exuded by large tech firms might be caught unaware by the ways in which their data is traded on the web.

My work will fit into this picture in two ways. Understanding profile reconstruction abilities and third party threat vectors will add a further dimension to the above assessments, pointing to potential future work on assessing user comfort with leaked data. Second, by surveying users on their outlooks, I hope to contribute directly to this discussion on customer expectations.

CHAPTER 5

APPROACH

In Chapter 4, I outlined the central research goals of this thesis and listed the questions I hope to investigate. In this chapter, I will describe my methodology. After providing an overview of my approach, I explain the processes I use to collect data. Then, I outline the techniques I use to analyze these data, organized along each central question. Finally, I briefly outline implementation details and challenges I faced.

Attempting to inspect ad targeting is an incredibly complex endeavor. Due to the intricacy of the advertisement placement process and the obfuscation of input-output links in bidding algorithms, 'ground-truth' data of any kind is hard to come by. Furthermore, because past works have successfully demonstrated the potential for important data to be leaked via advertisements, both companies and research groups are loath to release comprehensive datasets from real users. Finally, because of the competitive nature of the advertising landscape, methods used are

continually changing over time, meaning that techniques used successfully in the past may not work for long.

Combined, these realities meant that I needed to collect my own datasets and conduct novel analyses. Broadly, then, my approach was a three-step process. First, I used simulated, targeted web crawls (in the spirit of [54], [58]) to develop background intuitions on ad targeting methodologies (specifically as a proof of concept for question 3). Then, I used a Chrome extension to collect advertisements from and survey the privacy practices of users in real time.[13] Finally, I analyzed these datasets in light of each of my six research questions.

5.1 Data Collection

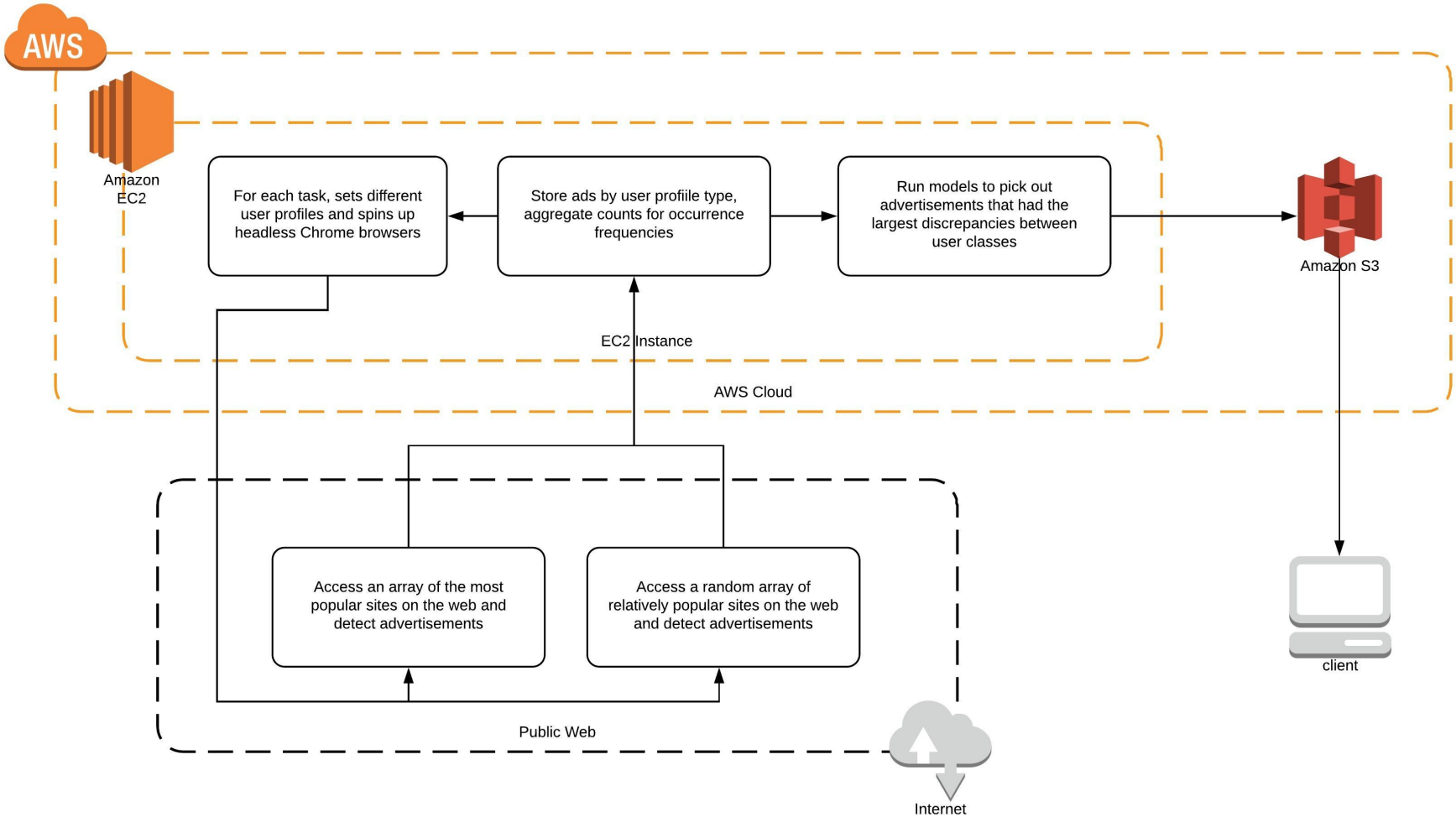
I collected two datasets: one from simulated users on orchestrated browsers, and one from live users. In building both data collection modules, I kept three common considerations in mind: I needed to write efficient and resilient code that could identify advertisements while respecting fundamental ethical principles. In the following sections, I describe my approach for collecting each dataset.

5.1.1 Orchestrated Data Collection

I first collected orchestrated data from simulated users in order to answer key question 3. Specifically, I hoped to generate a dataset of ads targeted uniquely to either women or men by simulating individuals of each gender accessing the same sites and observing the discrepancies in the ads that appeared. To do so, I leveraged the AdFisher code module, described and open sourced by [54], building in new functions wherever necessary. Broadly, I followed a three step approach.

[13] These methods were approved by the Princeton University Institutional Review Board (see appendix A).

Figure 5.1: *Orchestration module architecture diagram*



My modules repeatedly spun up Chrome browser instances and logged into Google as one of two pre-created profiles (one male, one female). Then, these browsers were told to access an array of the 500 most common sites on the web and a random but constant set of 300 less popular sites. Finally, these browsers hit a different collection of 500 sites and I collected all advertisements displayed on each.

Once I'd done that, I had generated a dataset consisting of ad impressions for each profile on each browser. To analyze these data, I updated AdFisher's built-in tools to join identical ads and aggregate counts by profile for ads hit. Finally, I conducted statistical analyses on these ad differences, noting whether certain ads were being disproportionately targeted to one of the two profiles. The list of the top 30 sites with the strongest statistical results for each profile became the gender-identifying sites I used when distinguishing gender in my own test set. Figure 5.1 provides an overview of this system.

While the existing AdFisher module provided much of the baseline logic for the orchestration process, there were still significant challenges I had to deal with when modifying the system for my own purposes. For one, AdFisher was brittle on the open-ended set of sites I visited, often erroring ungracefully as various expected site features didn't appear. I modified my own ad detection code (from my live user phase, below) to ensure that I was collecting parallel data through both approaches. Second, since when AdFisher was released, Google had changed their ad preference manager. My module couldn't simply set its gender at each run – ad preferences were only accessible to users who had signed in. Thus, I setup different Google accounts and prefilled them with basic characteristics (including gender), then added code to AdFisher that allowed browsers to successfully log in on Google before performing crawls.

5.1.2 Live User Data Collection

The second phase of my approach involved collecting data from live users. To do this, I needed to write a program that could do two things: automatically collect targeted ads served to users by Google or Facebook, and allow users to self-report their interests, demographic information, personal identifiers, and outlooks on privacy. On top of the standards mentioned earlier, this program also needed to satisfy stringent privacy and security standards and run effectively on a range of end-user devices.

To satisfy these constraints, I developed a Google Chrome extension that could perform both core functions effectively and safely. The extension consisted of a popup survey that users could fill at their convenience and a complex set of background scripts that silently processed advertisements on different pages. Beyond its user-facing features, it also interfaced with Chrome's built in storage functions (both local and synchronized) and API Gateways I setup on Amazon Web Services. Figure 5.3 provides an overarching picture of this architecture.

After a user consented to participation and installed the extension, they would see an icon in their main navigation bar for the "Leaking Ad Data Extension". Clicking on this icon brought up a form that users could fill out with information about themselves (see figure 5.2). This form had five sections: personal information, demographic information, privacy practices, user profiles, and privacy outlooks. The first asked for personally identifying information that would only be used to prune advertisements (to reduce the chance of inadvertent identification). These responses were never sent to my servers in plaintext; before hitting the API gateway endpoint, they were encrypted with a private key that existed only on the user's machine. The responses to each of the other sections were tied to a unique `userid` and sent to my `UserData` database. On each press of the save button, a custom Javascript trigger would process these answers and send them via a POST

request to the `/UserInfo` API Gateway endpoint I'd defined. Any answer fields that were updated were also sent to Chrome's `local-storage` (for PII) or `synced-storage` (for all else) functions. This way, each time a user re-opened the form, a different custom Javascript trigger could query all of these storage locations and pre-populate the form with a user's past responses.

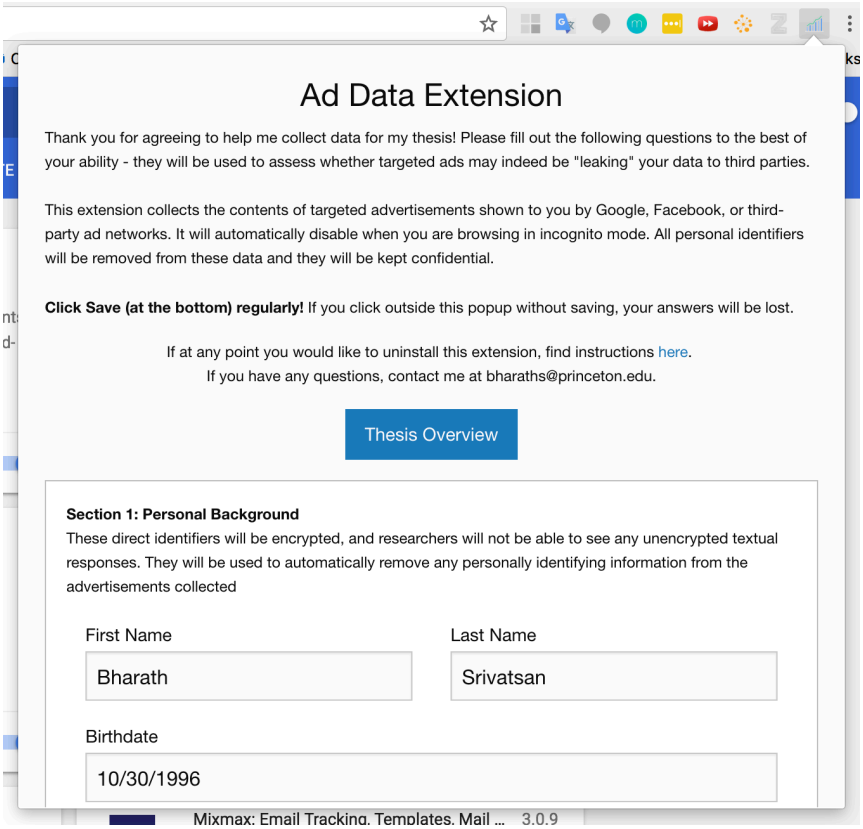
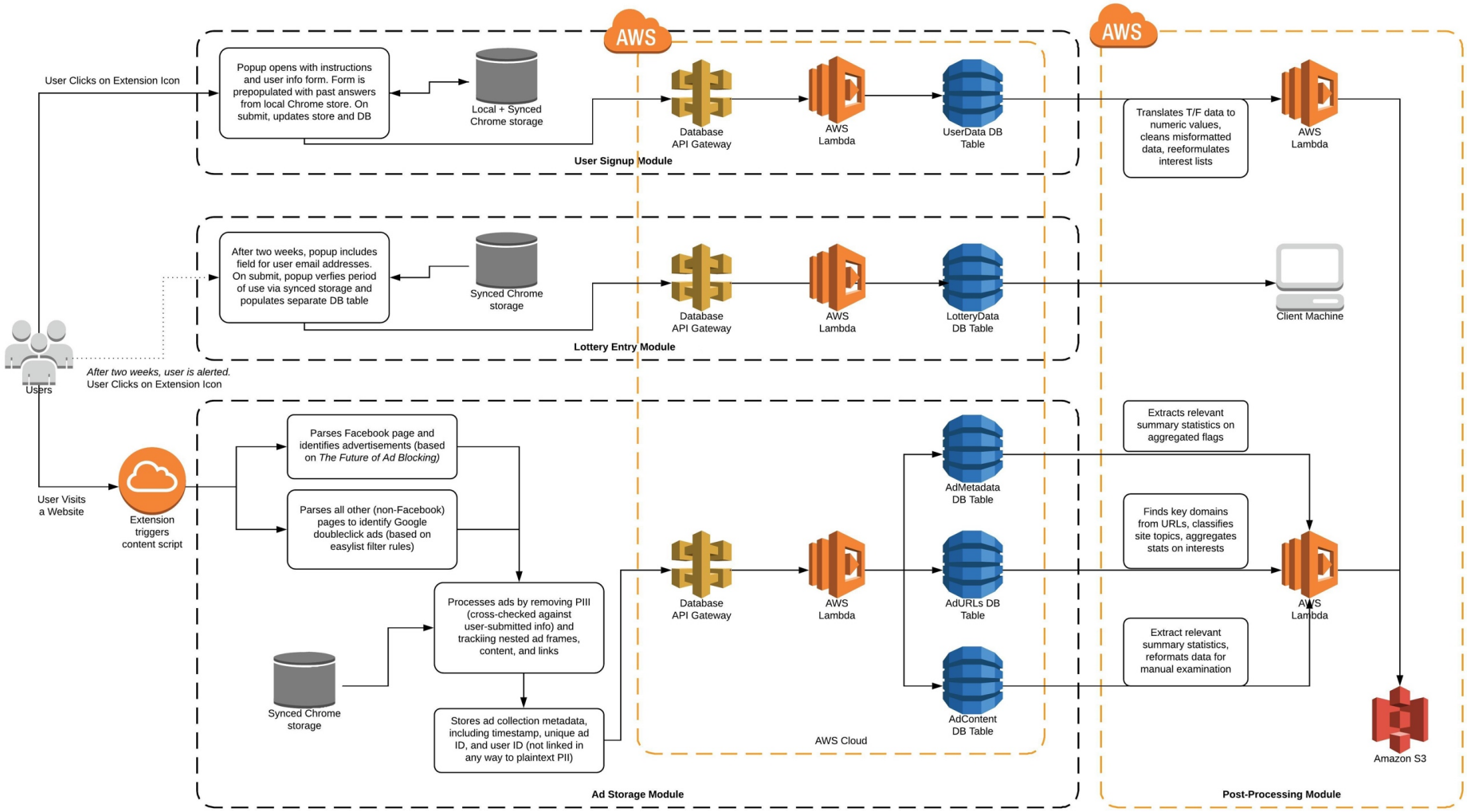


Figure 5.2: *Extension survey screenshot*

In order to ensure users only had to answer the questions they were comfortable with answering, none of the questions were mandatory. While this increased post-processing complexity, it allowed for partial completions and thereby encouraged more user engagement than might otherwise have occurred. From post-hoc manual inspection of these data, it seems that there weren't any individual users who took advantage of this feature by refusing to submit answers

Figure 5.3 *Live user module (extension) architecture diagram*



to an excessive number of questions.

When a new user saved the popup form for the first time, two things would occur. First, a `userid` would be set and content scripts to process ads would engage (more on this to follow). Second, an alarm would be set in the extension's background script to be triggered after two weeks. This alarm was linked to a Javascript function that revealed a lottery entry mechanism. After two weeks, when users re-opened the popup form, they'd see Section 0, which consisted of their `userids` (necessary for retrieving their personalized ad reports at the end of the study) and a lottery submission box for their email addresses. While users could theoretically reveal this box using their console (in order to submit their email addresses early), each submission was tied to their unique id and the time of their first save, allowing my server-side scripts to identify and reject such attempts. Furthermore, circumventing the two week alarm by revealing this box wouldn't reveal the user's id, since there'd only be placeholder text in the relevant space.

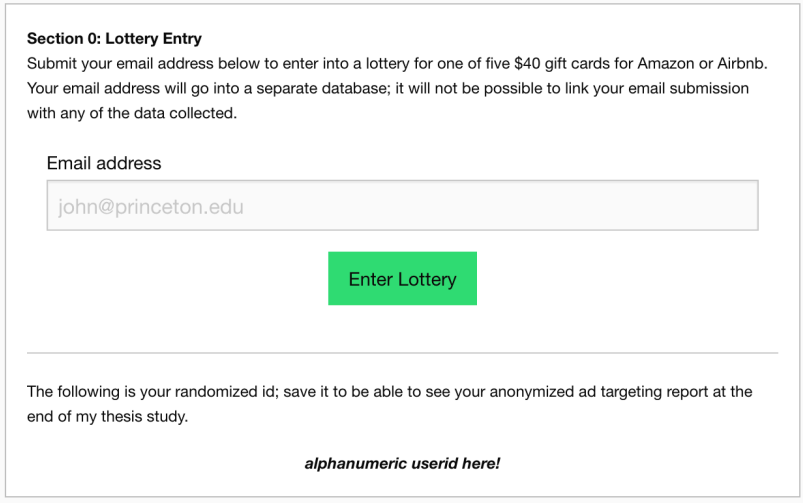


Figure 5.4: *Lottery entry mechanism screenshot*

The second major role of my extension was to identify advertisements across the web. Distinguishing advertisements from regular web content is no simple task,

and intentionally so. In order to complicate measures for adblockers, ad networks and ad exchanges continually refine their ad obfuscation techniques. For my purposes, I didn't need to collect every ad on a given page – over enough browsing sessions, I'd collect enough ads to conduct the analyses I'd planned on performing. I did, however, need to build a high-precision system that didn't inadvertently collect information outside of advertisements.

My ad processing module consisted of content scripts that were triggered on each new page load and a background script that served as a middleman for inter-frame communication. The content scripts did the main body of work, identifying and processing advertisements, while the background script coordinated messages and sent ads to my AWS API Gateway endpoint. Figure 5.6 illustrates the specific ways different code modules interacted with one another. Figure 5.5 demonstrates the difficulty of this task – across these sites, ads often appeared in remarkably different contexts.

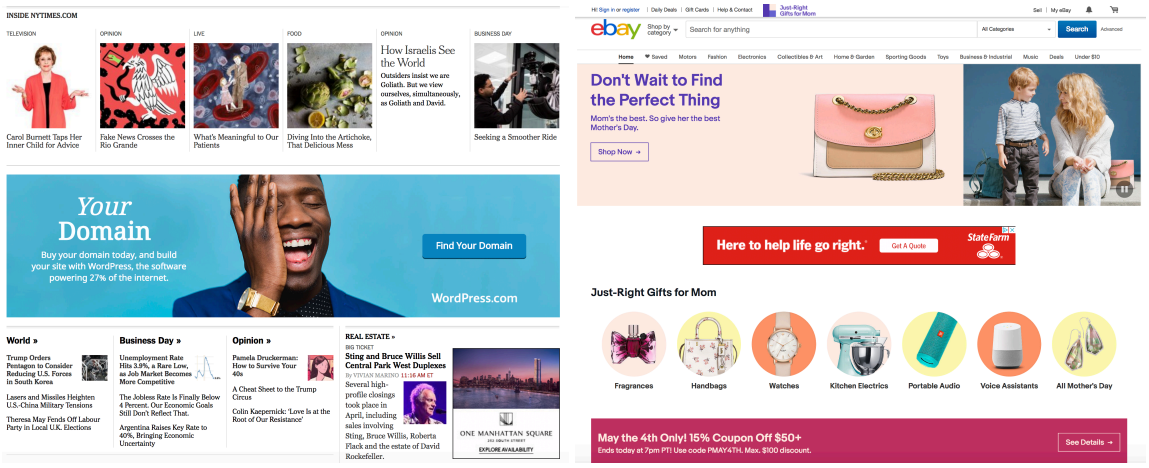


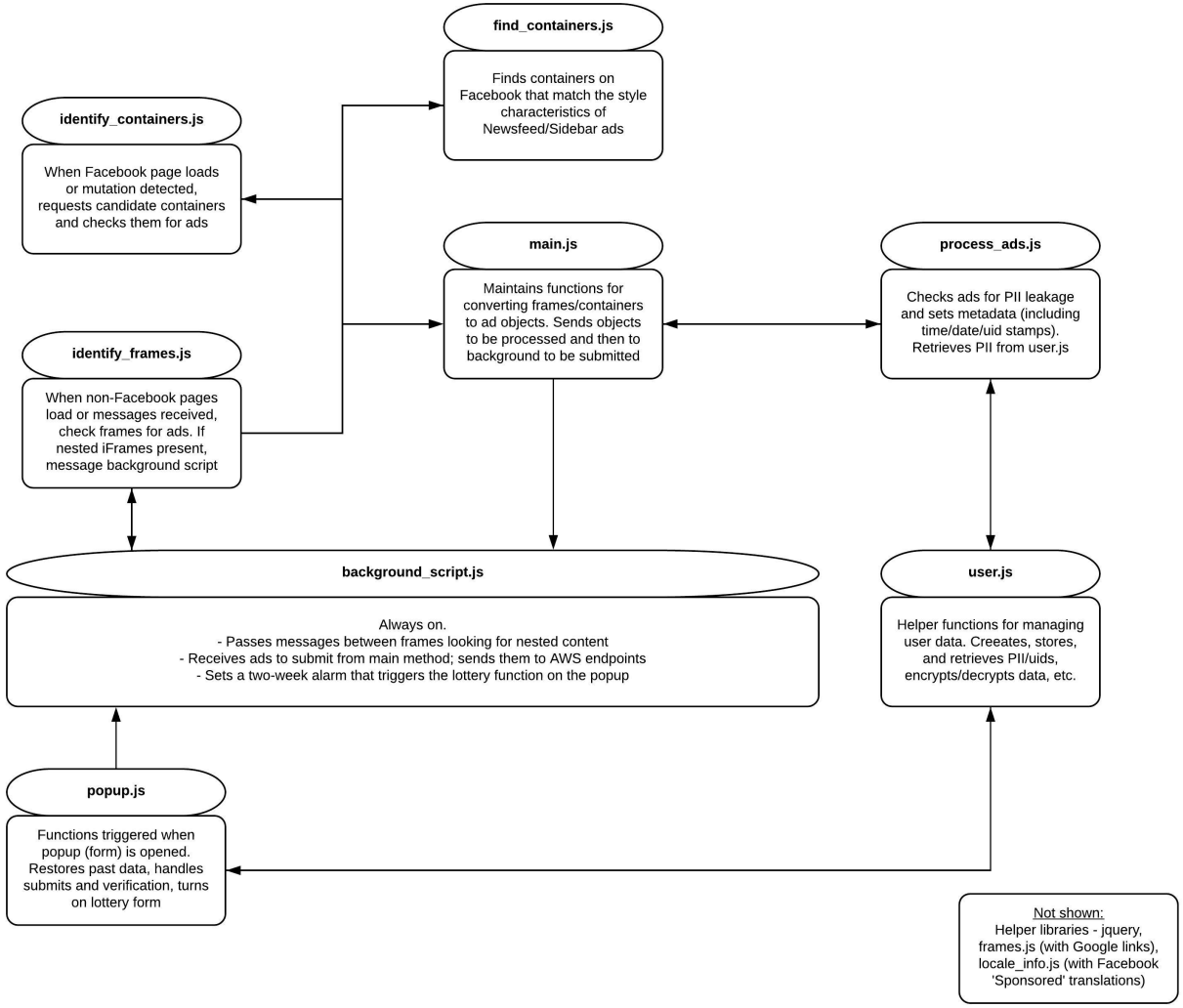
Figure 5.5: *Screenshots of ads in varied site contexts*

Given the vast differences in ad construction between ads served via Facebook's and Google's respective ad exchanges, I used different ad detection heuristics for each. My technique for finding ads on Facebook was modified from

the Perceptual AdBlocker, open sourced by [63] – like Storey et al., I took advantage of the fact that all Facebook ads explicitly note that they are 'Sponsored'. By recursively searching through containers for this text, I was able to identify both newsfeed and sidebar ads. I also borrowed Storey et al.'s technique of monitoring changes via a mutation observer, allowing me to discover ads placed as users scrolled further down their feeds or as sidebar ads changed (as opposed to only when users initially accessed the page). For Google, I developed my own technique based on the intuition behind many adblockers – Google-served ads would reveal themselves via exchange-specific text or formatting in their iFrames. I modified information from the Easylist set of advertising domains to include only those patterns linked to Google. Then, I recursively searched through the names, urls, and sources of iFrames for pattern matches. In order to bypass cross-site scripting protections, I coded this logic in content scripts that were injected into each frame. Since Facebook ads are served naturally (as opposed to in externally-sourced iFrames) I used separate content scripts to process Facebook at the top level. To find nested iFrames, I had the Google content scripts send messages to each other via the always-on background script.

After either of these units discovered an ad, my extension would process it for submission. Each advertisement was classified by type and class, allowing me to distinguish between newsfeed and sidebar ads, full-sized ads and single-pixel trackers, and Google ads served natively or on behalf of Facebook, Amazon, or third parties. Then, each ad was processed into three objects. `AdURLs` objects collected all of the linking and script URLs present in ads (or any nested iFrames). `AdContent` aggregated ad text and html. `AdMetadata` objects noted information on ad sources and parent domains. Finally, the `AdURLs` and `AdContent` objects were checked for PII – each found instance was removed and triggered a flag stored along with the metadata noting the type of PII that had been observed. These

Figure 5.6: *Extension code overview and diagram*



three objects were then stamped with userids, delivery times, and a unique object id, and sent to their respective API Gateway endpoints by my background script.

Once I'd built the Chrome extension, I needed to recruit participants for my study. To do so, I contacted friends and sent emails to listservs in the Princeton community advertising the study. In exchange for participating for two weeks, users would receive a shot at a lottery for one of five $40 Amazon/Airbnb gift cards and the promise of a personalized report detailing ways to better protect their data from advertisers. Users needed to be over 18, use Chrome for a significant portion of their browsing, and agree to disable/limit their adblocker for the duration of the study. When a potential participant indicated interest, I would send them my consent form, the extension files, and an instruction guide for installing the extension (see appendix A). Users were reminded that while personal identifiers would be removed, de-anonymization risks still existed.

Many of the design decisions I made in this phase of data collection were organized around the four non-functional considerations mentioned previously – efficiency, privacy, resilience, and ethicality. First, I needed to streamline operations and memory consumption to ensure that users wouldn't experience adverse effects as they browsed. To this end, I moved most of the expensive computational steps (accessing sites, categorizing topics, aggregating metrics) to post-processing scripts that could run on my AWS servers. The only operations that users' browsers ultimately performed were the absolutely mandatory ones - ad collection and PII pruning. The content script injection methodology was also chosen to maximize latency – scripts were injected in parallel and only after each page loaded, meaning that users wouldn't experience slowdowns in site rendering.

Given the potentially sensitive nature of the data I was collecting, I needed to ensure that my extension met stringent security standards for privacy protection. Broadly, this consisted of two mechanisms – protecting the privacy of

participants by cleaning data on client machines, and ensuring that the ad databases I was collecting were secured against attacks from malicious third parties. To guarantee that user data was kept confidential, I developed the PII-pruning mechanism described above to remove and flag leaked personal identifiers on the client's end (ie. before they arrived in my databases). Personal information and user data were strictly separated across tables, and identifiers were either encrypted (in the case of section 1 responses) or completely unlinked from user attributes like IP addresses or user ids (in the case of lottery entries). Protecting against external threat vectors required a more involved approach. First, I setup each AWS DynamoDB database to encrypt all data at rest, ensuring that even if attackers gained access to tables, they wouldn't be able to read their entries without the right keyfiles. Second, I setup data bindings and checks on both the client and server side, constraining avenues for potential SQL injection. Finally, I arranged the data collection components in a three step process. First, data hit an API gateway endpoint that could throttle aggressive user requests and that tracked IP addresses and request metadata (so that attackers could be blocked). This gateway channeled data to an AWS Lambda trigger that could process submissions to ensure that entries were clean. Only then were ads channeled to the relevant DynamoDB tables.

Since the extension was intended for a wide range of user devices and needed to work effectively on many different pages, it needed to be resilient to these sorts of changes. At a basic level, this meant that I needed to engage in significant amounts of testing across different machines and sites. In particular, I needed to ensure that errors wouldn't arise if users hadn't filled out fields on the forms or if they'd left various adblockers on. Beyond extensive manual testing, I split up the collected ad information into three ad objects (for content, metadata, and urls) and sent each separately. This allowed for more graceful erroring and higher success

rates for POST requests (since the likelihood of particularly large requests being throttled was drastically reduced).

Finally, I needed to take into account two major ethical considerations. First, as with any user study, I took pains to ensure that users were well aware of the risks and rewards associated with the study. Along these lines, I required users to fill out a comprehensive consent form (approved by the IRB) and included an explanatory document overviewing my thesis in my extension (see appendix A). Second, I needed to ensure that I wasn't breaking any web-based ethical standards through my data collection. The main risk here was violating norms on click fraud in advertisements. By post-processing ad urls in batches (across users) and by only clicking on ad urls when natural language approaches to find destination urls in full links were unsuccessful, I minimized the number of ad-clicks performed.

5.2 Analysis

After collecting each dataset, I first investigated how often Personally Identifying Information (PII) was leaked through advertisements. Though the US Department of Commerce's National Institute of Standards and Technology (NIST) releases guidelines on PII, the body does not explicitly articulate what specific types of data count as PII. Other researchers have suggested moving beyond the PII framework entirely and nuancing the conversation into a wider range of user data categories [91]. Indeed, techniques like fingerprinting mean that seemingly random bits of information can be used to identify individuals. For the purposes of this work, however, I define four kinds of personal identifiers to track in ads – names (either first or last), birthdates, current locations (cities, states, or countries), and home locations (likewise). Individual sites may be leaking further identifiers like userids and email addresses to third parties or advertisers, but given the difficulty of tracking such a diverse range of possible disclosures, I focused on

these four features. After users self-reported each of these characteristics, each ad was pruned and flagged for PII presence on the client side. To reach general conclusions, I wrote a script that aggregated flag counts by user for each type of leakage.

I then investigated research question 2, on insecure, sensitive, or malicious ads. I defined insecure ads as those that used HTTP (as opposed to HTTPS). There were many ways that ads could do so – they could be sourced from HTTP sites, implant HTTP-sourced assets (like images or scripts), or link to HTTP targets. Unencrypted traffic along these lines could open users to potential threats (if cookies or flags were transmitted in plaintext, for instance), so I counted and flagged HTTP usage across all urls embedded in ads. I aggregated these counts in two ways – first by counting the proportion of ads that had at least one HTTP link, and then by counting the total number of HTTP links seen by each user. To detect sensitive and malicious links in ads, I needed to categorize the content of the sites they linked to. For this, I relied on a script that extracted target links from ads (either through unpacking ad urls or by simulating clicks) and then passed them on to the WebShrinker Category API.[14] I manually compiled category lists (see table 5.1) that were potentially sensitive or malicious, and crosschecked the returned site categorizations against this list. I counted an ad as sensitive or malicious if it included at least one category in the corresponding list, and aggregated counts of such ads by user. In doing so, I acknowledge two flaws: first, some types of sensitive disclosure are potentially 'worse' than others – for some users, it would be worse to find ads about sexual orientation than health conditions. However, given the immense variance in the impact of sensitive disclosure risk by user, I rely on an aggregate statistic instead rather than making blanket claims

[14] Building a categorizer from scratch for an open-ended set of possible websites would have been an immense task prone to many errors; of the categorization services I investigated, WebShrinker was the most affordable option that had an adequate coverage of sites on the web.

Type	IAB Category	Subcategories/Topics
Illegal Content	IAB 26 IAB 19	Illegal Content/Wares Illegal Drugs/Paraphernalia Hacking/Cracking
Non-Standard Content	IAB 25	Adult Content Profane Content Hate Content
Personal	IAB 14 IAB 23	Religion and Spirituality Dating/Personals LGBTQ+ Ethnic Content
Health and Wellness	IAB 7 IAB 6	Panic/Anxiety Disorders Abuse Support Women's Health Pregnancy

Table 5.1: *Selected sensitive IAB categories*

about which sensitive categories are worst. Second, because of the difficulty of website categorization generally I cannot guarantee that WebShrinker's categorizations were all accurate, despite my efforts to verify some site responses.

To understand whether demographic categories could be reconstructed from targeted ads (question 3), I used two approaches. The first leveraged the orchestration data generated from my simulated web crawls. I'd collected sites whose ads appeared disproportionately to either men or women, suggesting that they were targeted by gender. I then iterated through the domains that targeted ads to each live user, noting the number of hits from the gendered sets above. This yielded a count of male-targeted impressions and female-targeted impressions for each user. To assess the skew of these gendered ads, I calculated the relative proportion of male or female-targeted impressions to the total number of gender-targeted impressions, and proposed for each user the gender corresponding to the

higher proportion. Finally, to reduce the rate of false positives, I experimentally developed heuristics to classify only those users with sufficiently disproportionate targeting (higher relative skews) over sufficiently large ad hit rates (more gendered impressions). In this way, I removed classifications in marginal cases (eg. only 55% female-targeted ads) or that could be ascribed to random volatility (eg. only 5 gendered ads shown). Ultimately, I classified users with over an 80% skew and over 30 gendered ad impressions. This analysis was performed separately for both Facebook and Google ads.

The second approach I used involved manually inspecting ad targeting graphs for patterns corresponding to various demographic features. I began by compiling for each user the top domains and top IAB categories from which they received ads. Then, I created two types of graphs, both with users represented as nodes: one connecting users that shared at least one top domain and the other connecting users that shared at least one top IAB category. The number of domains or IAB categories to use was derived experimentally – using the top 5 for each user generated graphs that were too densely connected to easily examine for patterns, while using just the top category for each user created sparsely connected graphs with many isolated nodes. I settled on using the top two domains/categories for each user, as this created graphs with distinct and meaningful clusters. This process was repeated for both Facebook and Google ads/categories, yielding four canonical graph structures in total. Finally, I colored the nodes in each graph by various demographic characteristics and noted patterns in targeting behavior. I began by coloring users by gender and/or race to observe how ad targeting differed across each category. Then, I turned to more nuanced Facebook categories, coloring green, for example, all users who'd been designated by Facebook as 'away from hometown' or 'away from family' and red all others. For a full list of categories used, see table 5.2.

Category Description	Values
Gender	Male | Female | Other
Race	White | Asian | Other
Displaced	"Away from hometown" / "Away from family"
Traveler	"Frequent Travelers" / "Frequent international travelers" / "Close friends of expats"
Liberal	"US politics (very liberal)" / "US politics (liberal)"
Shopper	"Engaged Shoppers"

Table 5.2: *Demographic features/user attributes tested for*

To evaluate interest classification for key question 4, I applied a three step process for each user, inspired by [55]. First, I counted the number of distinct domains from each IAB category that served a given user ads. This was then used to generate a ranking of categories from which a user had received the largest number of ads. I used this distinct-domain approach (rather than a raw impressions count for each category) in order to distinguish between genuine interests and retargeted ads; some sites would serve many ads to users who had previously visited them, potentially throwing off my interest classifications. Finally, I compared these categories to the Google interest lists that users had self-reported. Unfortunately, Google's interest categories did not match up perfectly with the IAB set; in order to correct for this I manually developed a matching between the two lists. To numerically evaluate reclassification success, I took the top 1, 3, 5, 10, and 20 categories and computed precision, recall, and F-Score figures for each against the ground-truth Google sets. Assessing success against Facebook's interest lists was significantly harder given that Facebook used an open-ended set of brands and topics to represent user interests (see figure 6.22). For these interests, I manually compared category lists for some users to identify particularly interesting features.

Finally, I unpacked the survey responses I'd received to answer questions 5 and 6. I first computed basic summary statistics on numeric responses – averages, ranges, and standard deviations. I then plotted responses to visually represent these data. Finally, I ran t-tests and correlation calculations to examine differences across categories and links between them.

5.3 Implementation

Due to its ubiquity and the wealth of resources built around its extension ecosystem, Google Chrome was an easy choice as a baseline platform for my live user data collection. For consistency's sake, Chrome was also the browser of choice used in my simulated user experiments. While the extension itself was run on user browsers, I made use of both a local virtual environment and a range of AWS services for specific modules. AWS API Gateway was used to set up data collection endpoints, Lambda for data pre- and post-processing, DynamoDB for storing ad and user data in SQL form, S3 for holding downloadable, cleaned csv files, and CloudWatch for keeping and tracking logs. An EC2 instance, chosen for its high compute performance, was used in the orchestration phase and for some of my processing scripts.

The majority of my code was written in Javascript, HTML, and Python. Since the extension relied on content scripts and webforms built for Chrome, I needed to write it in Javascript and HTML. Meanwhile, the Lambda triggers and post-processing code required efficient performance on potentially massive datasets, a task far better suited to Python. The orchestration module I implemented was based on Adfisher, which was itself implemented in Python [54].

5.4 Challenges

In conducting this research, I ran into a range of challenges, some particularly difficult to resolve. In this section, I identify three of the largest I faced.

5.4.1 Ad Identification

The largest implementation challenge I tackled was in trying to build my ad identification module for live user data collection. In both my Facebook and Google data collection scripts, unique behaviors implemented on both sites' exchanges (presumably in some cases to discourage adblockers) complicated my collection process drastically.

Google's ads were taken from iFrames present on third party sites across the web. To identify them, my content scripts checked for Google-related information in iFrame tags or content. This approach had proven quite effective for adblockers, since they could simply block all top-level Google-distributed iFrames. For my purposes, though, I discovered that this approach had a massive flaw. In some ad frames, the actual ad content was nested in inner iFrames, often from completely different sources. Due to the same origin policy, a content script loaded on one frame can't read data from nested frames from different sources. This meant that if the top-level Google-annotated iFrame didn't have relevant ad content or links, the ad objects generated would be useless. Meanwhile, the inner iFrames (those that contained the actual ad information I needed) were often not tagged with Google-related exchange information, meaning that my content scripts wouldn't be able to identify them. To complicate things further, this nesting could be quite complex – in some sites, ad content was hidden under five layers of iFrames, each from a different source.

To resolve this issue, I rebuilt my ad identification module from the ground up. I first found a way to uniquely identify frames by their positions on the global DOM tree. Whenever a Google-identified ad frame was detected, it would send a request to the background script with the tree positions of all iFrames nested immediately beneath it. This message would be passed on to all iFrames, and the content scripts of iFrames at those requested positions would respond with an ad object for their frame (regardless of whether they were tagged by Google). This process happened recursively; if a nested frame had other iFrames beneath it, it would send a similar background script request with its own desire for inner object contents. In this way, each Google frame included all iFrame content nested below it - no matter how deep these trees went. One additional complication that arose from this method was that frames were sometimes dynamically created and destroyed, meaning that frames might wait forever for responses from nested objects that had already disappeared. Thus, rather than having frames wait for all nested content, I implemented a timeout mechanism that sent along ad content even if all responses hadn't yet been received.

Meanwhile, on Facebook, the ad identification mechanism developed by [63] searched for the word "Sponsored" in ad containers. This mechanism already took care of some edge cases – for example, it translated the word into different languages based on locale and circumvented the fact that Facebook sometimes split up this text into many different divs. When I initially ran this code, however, I found that my module was only non-deterministically finding ads; anecdotally, I was capturing 15-20% of advertisements. After a long process of manual inspection, I discovered that Facebook was occasionally injecting random sequences of the letter "S" in hidden divs in the middle of the word "Sponsored." This meant that the original implementation would sometimes see containers tagged with "SpSSSonsoSSresssd" or the like and ignore them. To fix this, I used a regular

expression search instead that could circumvent this insertion. While this is by no means a long-term solution (Facebook could simply insert other letters instead), it was sufficient for significantly improving my own detection metrics.

Finally, in my original implementation, after ad objects were compiled and processed, the top-level content script would attempt to send them to the API Gateway endpoint. I soon realized, however, that only a small fraction of the ads I was detecting and processing were being collected. This was because in the time it took the content script to prepare and send each POST request, the frames or sites themselves would often be closed, updated, or destroyed. To resolve this issue, I had content scripts forward ads found to the background script (which was always-on and could asynchronously send many of these requests) for POST-ing.

5.4.2 Data Collection

In designing my live user study, I had to make content collection choices carefully. Since potential respondents likely cared deeply about their privacy, I needed to navigate a tradeoff between the volume of data collected on individual users and the willingness of new users to participate. One of the main ways this came to bear was with respect to user browsing histories and cookie placements. Understanding a user's past activities would have allowed me to analyze retargeted ads and improve my interest categorization methods. Using Chrome WebAPIs for user histories and cookie databases, I would easily have been able to include these data in my collection. However, from preliminary discussions with potential study participants, I realized that site browsing habits were of particular interest to users wanting to protect their privacy from researchers. As such, I ultimately did not petition to collect either type of data when designing my experiment and proposing it to the IRB.

Even after I'd gotten approval for my study, I still needed to make individual methodological decisions with user security and convenience in mind. For example, though I'd been authorized to collect screenshots of advertisements, I soon realized that doing so would be difficult. For one, the actual image collection and encoding mechanism built into Chrome was particularly inefficient – simply including this feature in my extension dramatically slowed down ad object collection (and took up a substantial amount of space). The nested frame issue identified in §5.4.1 also meant that screenshots would need to be taken at each level, further compounding latency and storage issues. Second, it would be far tougher to remove personal identifiers from images, meaning that the risk of de-identification would drastically increase. Though having image content might have assisted in some parts of my analysis, I also noted the fact that many ads appeared in video form or contained images that weren't personalized to individual users. This meant that the additional information I'd get from collecting images likely wouldn't have outweighed the potential risks.

5.4.3 Confound Controls

Finally, in line with concerns outlined in [38], minimizing the effects of confounding variables was a particularly difficult task. I needed to be careful not to draw demographic conclusions from ad observations based instead on location or automated A/B testing, for example.

I attempted to mitigate these possibilities in two ways. First, I collected orchestration data over a short time horizon and from browsers simulated at the same location. Beyond the different Google accounts logged in to, these browsers were setup identically and followed the same trail of sites. In this way, I minimized the potential for confounding variables appearing in a non-controlled way. Though homogenizing live user ads was far trickier, some natural features of my dataset

reduced these concerns. Perhaps most significantly, most users were from the Princeton, NJ area and browsed naturally over a two-week period. This reduced the effects of location, and while A/B tests may have influenced how ads were tailored and scoped, they likely did not affect whether ads appeared to this small set of users over this small amount of time in any consistent manner. I noted those users that were outliers by age, education level, or location, and wrote a separate script that flagged whether they were outliers in my analysis metrics – for instance, in their sensitive ad hit rates or their interest reclassification accuracies – but did not find any instances of this.

More broadly, I qualify repeatedly that I do not hope to reach definitive answers about how ads are targeted. In what follows, I use targeted ads in a variety of ways – to infer user demographics, to guess at sensitive characteristics, etc. – but do not aim to 'prove' that users received particular ads because of particular features.

CHAPTER 6

FINDINGS & DISCUSSION

In this chapter, I review the findings of my work. I begin by providing an overview of the datasets I collected – from both orchestrated web browsers and live users. I then turn to describing my results, walking through what I learned about each of the six research questions defined earlier. Finally, I pick out two illustrative case studies to analyze in more depth. Throughout the chapter, I emphasize particularly important statistics.[15]

6.1 Datasets

As outlined in §5.1, I collected data from both simulated and live users. The former method involved setting up an EC2 machine to scrape ads associated with

[15] Note: as mentioned in Chapter 5, the survey questions included in my live user study were not mandatory. This means that respondent/analysis counts across questions may not add up to the same totals.

Male	Female	Ungendered
reddit.com	revolve.com	amazon.com
espn.com	kohls.com	airbnb.com
dollar.com	progressive.com	reddit.com
express.co.uk	vox.com	open.spotify.com
tableau.com	lulus.com	nytimes.com

Table 6.1: *Top domains serving ads to simulated users by gender*

different demographic profiles, while the latter gathered both advertisements and survey responses from real users.

6.1.1 Orchestrated Data Collection

In my simulated data collection module, I ran 6 browsers (two male, two female, and two with no profile) and collected **24,837 advertisements** from repeated hits on 500 sites. These ads resolved to a set of just over 1,300 unique sites, with some serving ads at significantly higher rates than others. I ranked each of the sites served to the male profile by noting statistical discrepancies against impression rates for the female and non-logged in profiles (and vice versa for the female list). I then took the superset of the top 30 sites from each ranking to serve as the list of canonical male/female ads to search for in my demographic analysis. See table 6.1 for the top domains from each list.

6.1.2 Live User Data Collection

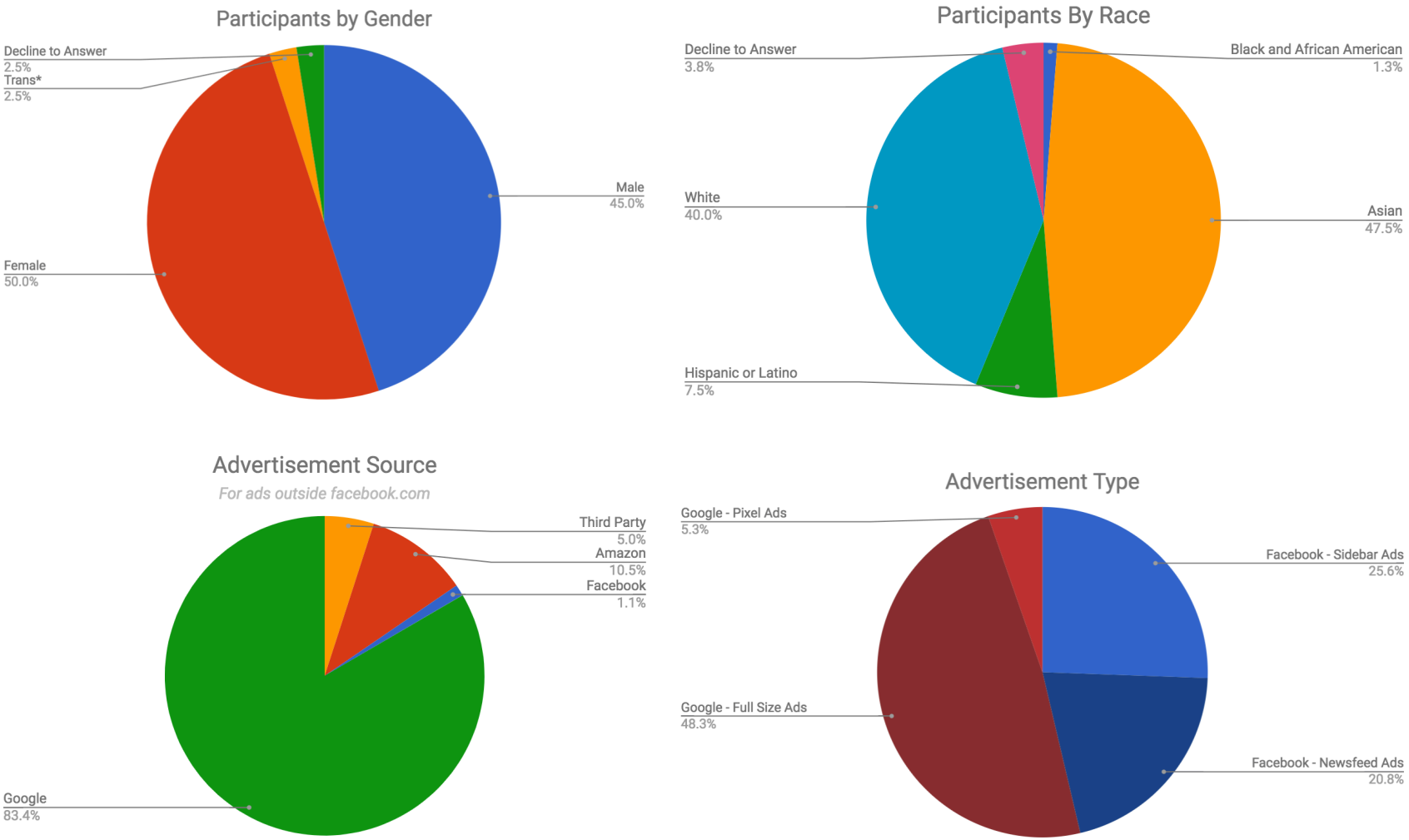
Over the course of two weeks, I collected over **60,825 advertisements** from **80 participants**, making this work (to my knowledge) the largest live-user study of targeted advertisements to date. Each of these users were also asked **33 questions**, contributing further information for analysis (see appendix B).

Participants were well distributed by gender – 36 identified as male, 40 as female, 2 as trans*, and 2 declined to answer (see figure 6.1). However, as a result of natural sample selection difficulties in a college environment, users were quite clustered by age and educational background. The average age across all users was 21.1; 73 of 80 participants were college-aged (between 18 and 22).[16] Meanwhile, 89% of respondents indicated they had or were pursuing an undergraduate degree. With respect to race, we found large numbers of white and Asian participants (see figure 6.2). No American-Indians, Alaskan Natives, Native Hawaiians, or Other Pacific Islanders participated in our survey.

I found a good diversity in the advertisement data I collected (see figures 6.3 and 6.4). I collected **760 ads on average per user**, but found a massive standard deviation in collection rates – from as few as four ads for one user to as many as 1,884 ads for another. These ads came almost equally from my **Facebook and Google identification modules** (**46.3%** and **53.7%** of all ads, respectively). Of ads collected from sites outside of facebook.com, the majority were directly distributed via a Google-run ad exchange or network.

In total, I found ads from **4,743 different domains.** Simply examining domain counts yielded a few interesting conclusions. Perhaps unsurprisingly, Amazon served the largest number of ads on both Google and Facebook (by a large margin), though this count included small-scale businesses who linked their ads to Amazon pages. Many popular ads were from sites aimed at college students: thetab.com is a youth new site with offshoots at various colleges, ratemyprofessors.com is a rating site for college professors, and storagesquad.com is a popular summer storage option for students at Princeton. There was a substantial divergence between domains that advertised on Google and Facebook – many sites with narrower target

[16] After removing two outliers (>3 standard deviations from the mean), the average age dropped to 20.5. A plurality of participants were 21.



Clockwise, from top left: Figures 6.1: *Participant breakdown by gender;* 6.2: *Participant breakdown by race;* 6.3: *Advertisement breakdown by source;* 6.4: *Advertisement breakdown by type*

Google		Facebook	
Domain	*Count*	*Domain*	*Count*
amazon.com	2669	amazon.com	2815
revolve.com	875	greenhouse.io	2301
reddit.com	772	storagesquad.com	1723
ratemyprofessor.com	637	peiwei.com	1011
cronometer.com	614	nationalguard.com	621

Table 6.2: *Top domains serving ads to live users (by platform)*

audiences (eateries like peiwei.com, shopping sites like lulus.com and zaful.com, and cultural sites like birthrightisrael.com and avodah.net) advertised primarily on Facebook, while larger sites like reddit.com, nytimes.com, or spotify.com focused on Google. Table 6.2 lists the most common advertisers on each platform.

6.2 Results

Before discussing how these data pertain to each of the six research questions, I first note the sampling issues associated with this collection. These participants are by no means a representative collection of some well-defined class. Compared to Facebook or Google users on the whole, they likely bias younger and more educated; compared to Princeton University students they are disproportionately white or Asian. The recruitment methodology I used relied on opt-in interest from friends and fellow students. This meant that respondents were particularly likely to have a preexisting interest in advertising and Facebook/Google data protections, and may therefore engage in different privacy practices than a more general population sample. Similarly, the self-driven installation procedure likely selected for students that were particularly technologically literate (which may also have shaped their views).

Given these issues, I will begin by warning that these results are by no means prescriptive. I make no guarantees that my findings will generalize to different user classes, or that they can be used to inform how canonical users ought behave on the web. In what follows, I focus on illuminating intra-sample differentiations as opposed to making claims about generally-held beliefs or practices. For instance, I note in section §6.2.5 the effects of adblockers on perceived classification quality, but caution against using the raw adblocking averages to make claims about how many users in the real world use such tools. As and when specific methodological limitations arise, I note them in the sections below. With that said, however, I believe these data can give us valuable information – even if just as a proof of concept for the techniques I utilize. The leakage of PII or sensitive information, for example, is worrisome even if only limited to similar samples of users, as is an ability to recreate user demographic or interest profiles from ads alone.

6.2.1 Personal Identifiers

Using my flag-setting mechanism, I discovered that personal identifiers were leaked through both Facebook and Google networks. Most strikingly, just **over 1% of ads leaked users' locations** via their content or links (see figure 6.5). Even though the rates of name and home location leakage were lower, the volume of ads collected means that there were many such occurrences. No instances of birthday leakage were found, though the demographic reconstruction techniques explored in §6.2.3 may be effective at predicting user age nevertheless.

These figures are surprising. Both Google and Facebook attempt to prevent PII-leakage by setting firm guidelines on permissible information in ads. This work, though, suggests that their pruning mechanisms may not be as effective as previously thought. Even if we set aside Google's and Facebook's own promises on

the matter, PII-leakage is particularly important to monitor and curtail. As I will go on to demonstrate, targeted ads are revealing of a range of private user characteristics – leaked PII would open up the possibility of tying this picture to an individual in the real-world.

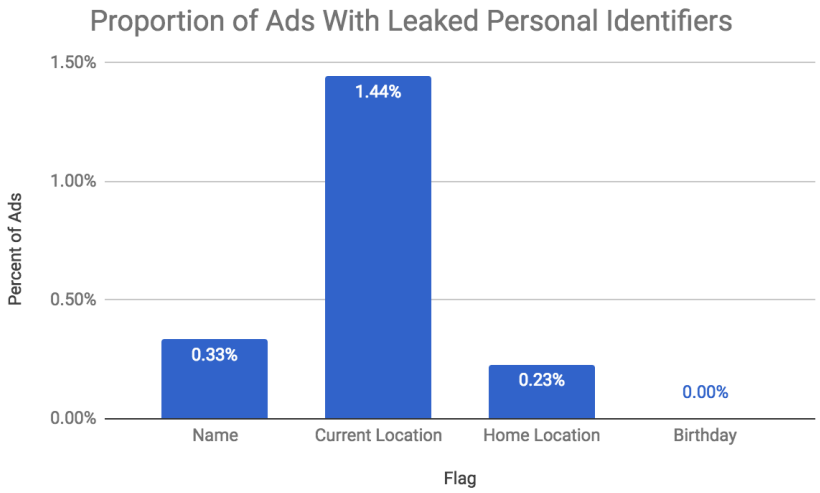


Figure 6.5: *PII leakage rates by type*

Furthermore, these algorithmic measures do not reveal the whole story. In concert with other information, even seemingly innocuous ads can de-anonymize individuals. For instance, when manually looking over ad sources, I noticed that a decent number of ads to one user were from sas.edu – my high school. Given that there likely weren't many participants in my study who would receive advertisements from a high school in Singapore, I could reasonably infer that this profile was my own.

6.2.2 Sensitive Sites

I then attempted to evaluate whether ads contained links to sensitive, malicious, or insecure content. I first found that ads often contained HTTP urls

either in ad sources, script sources, or ad target link redirects. Just as interestingly, I found that proportions of HTTP ad links differed sharply by user and service. On average, **28.97% of ads on Facebook** and **13.37% of ads from Google** contained at least one HTTP link or redirect. As figures 6.7 and 6.8 demonstrate, Facebook's distribution of HTTP link proportions by participant was particularly heavily skewed, culminating in a high outlier at 7.82 HTTP links, redirects, or sources per ad for one unfortunate user.

Unpacking sensitive ad content revealed similarly compelling insights. Across all users, **5.91% of ads on Facebook** and **6.10% of ads on Google** originated from sites tagged with at least one sensitive ad category. Even more so than HTTP site hits, however, sensitive ad proportions were severely skewed by user. Median sensitive site proportions were 3.66% and 2.85% for Google and Facebook, but proportions ranged from 0 to 26.16% and 0 to 42.75%, respectively. Also of note was the lack of a correlation between Facebook and Google sensitive ad proportions (see scatterplot in figure 6.6).

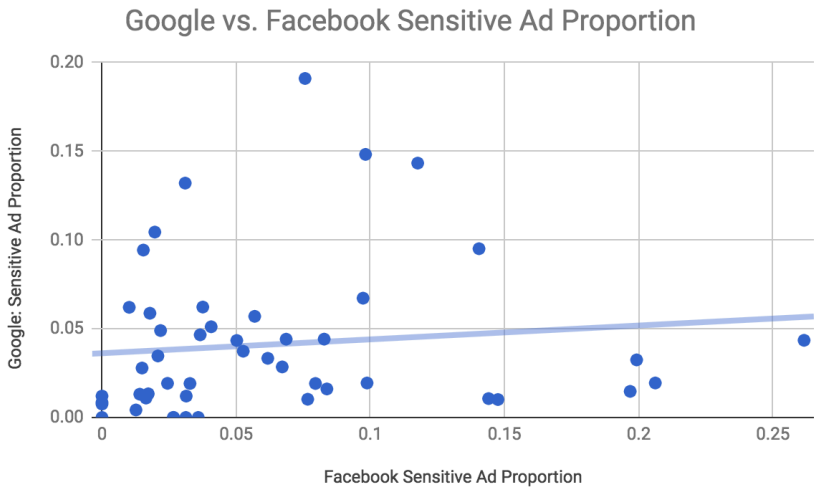
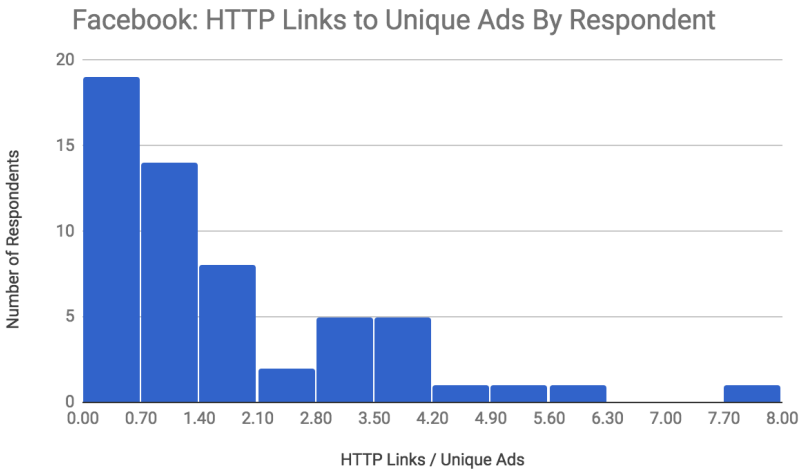
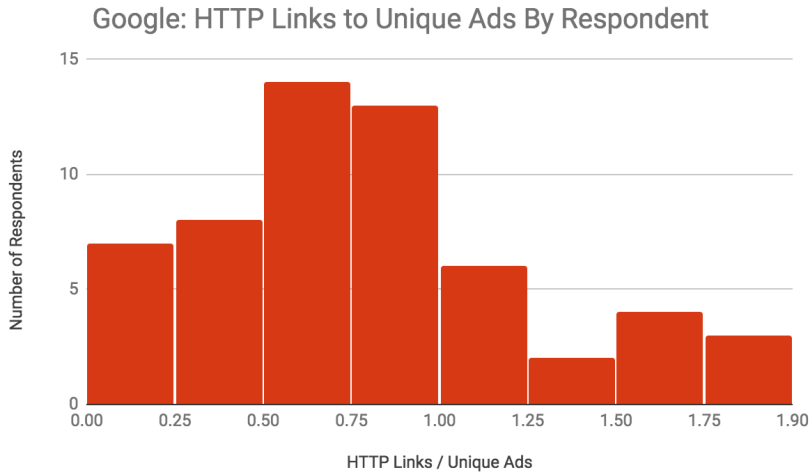
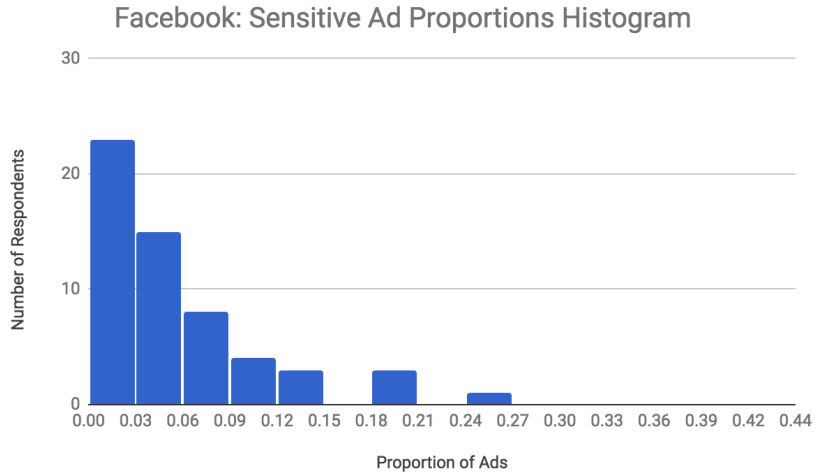
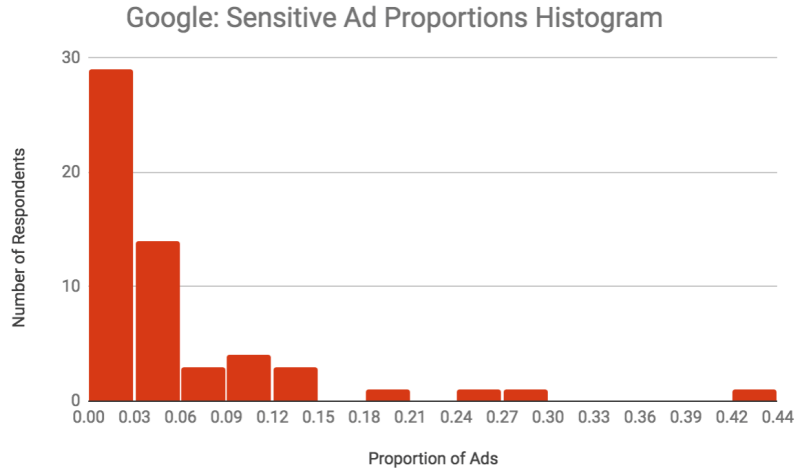


Figure 6.6: *Sensitive ad proportions for users of Google and Facebook*
N.B.: *This plot only includes the 48 users with over 50 ads from both platforms*









Clockwise, from top left: Figures 6.7: *HTTP link presence on Facebook ads;* 6.8: *HTTP link presence on Google ads;* 6.9: *Sensitive ad proportions on Facebook;* 6.10: *Sensitive ad proportions on Google*

Once again, manual observation revealed new dimensions to this picture. Some individuals saw ads for sites that were very sensitive but local (ie. low-traffic) and so hadn't been classified by WebShrinker. For instance, a Princeton-specific suicide helpline took out support ads that appeared for two users in my study. This type of ad is *incredibly* revealing and highly concerning. Others saw ads that were revealing in their content. One user received multiple ads from a popular real estate site with directives like "Complete Your Rental Now" and "Finish Signing Your Lease"; these specialized instructions indicated that this person was close to moving.

I found **no instances of malicious sites being linked** to via advertisements. This is a pleasant but unsurprising finding – it's unclear why such sites would spend money on digital advertising via Facebook and Google in the first place, and it's likely that these big networks have built in blocks against them. The HTTP and sensitive site hit rates I found, however, are worrying. As discussed in §4.1.3, unencrypted cookies and web data sent to advertisements could open avenues for exploitation. Further, the possibility of partial ad observation means that sensitive content in ads could end up posing real-world problems for recipients. The skewed distribution of such content indicates that it is targeted (a finding in line with past studies like [54]) which only heightens the risk that such individuals would face these adverse effects.

6.2.3 Demographic Information

The gender classification module I described in §5.2 made gender predictions for 47 participants on either Facebook or Google ad sets. 29 of the predictions were for women and 18 for men; **all 47 predictions were accurate**. Perhaps more surprisingly, 23 of the predictions from Google data were for users whose genders hadn't been accurately classified by Google. These results suggest both that

targeted ads are revealing of a recipient's gender, and that the information latent in ads can go beyond what is present in ad preference managers. This finding has two important implications. First, it demonstrates, as a proof of concept, that despite the possible confounds present in ads shown to live users, gender-based targeting can still be used to reclassify user attributes. Second, it implies that the information available to the public on advertising targeting may not be comprehensive – Google accurately classifies gender for only 52% of participants (see figure 6.11), but ads shown to a larger subset seemed to be indicative of gender. This suggests that though Google and Facebook may disclose the attributes they've predicted for users, third parties may be using more detailed pictures to make ad targeting decisions.

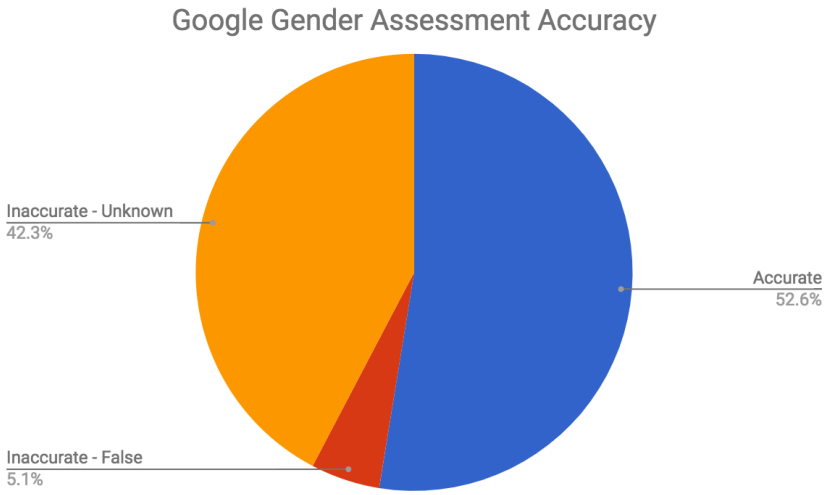


Figure 6.11: *Google gender assessment accuracy*

I include a few of the results of my graph clustering module in figures 6.12 to 6.15. In figure 6.12, I color users by gender and join them by the top categories of ads they received on Facebook. While the cluster of users whose top two ad categories included food and drink ads were mixed by gender, the cluster that disproportionately received ads on shopping mainly consisted of women.

Meanwhile, the users whose top ad categories included tech and computing were largely male. Figure 6.13 displays an even more worrying finding – by joining users based on the top *domains* serving them ads, I found that although all of the users that disproportionately received ads from niche shopping sites were female, the majority of users whose top advertisers included greenhouse.io, a recruiting platform, were male. Though we cannot assume from these graphs alone that these ads were targeted based on gender, these figures demonstrate in concert that gender can affect the makeup of ads that a user receives and that it sometimes does so along traditionally sexist lines. Figure 6.14 goes on to show that this effect isn't isolated to Facebook. Almost all users who disproportionately received Google ads linking to reddit.com were Asian men, while disproportionate recipients of Buzzfeed ads were Asian women and disproportionate recipients of revolve.com clothing ads were white women.[17]

To validate that these results could not simply be explained by disproportionately high ad counts from other categories, I compared the number of ads received from the categories and domains mentioned above for users of each gender. I found **statistically significant differences** by gender in the number of shopping ads, tech and computing ads, and ads for/from greenhouse.io, reddit.com, and revolve.com shown to users.

Coloring users by Facebook attributes was similarly revealing. As indicated in figure 6.15, travelers disproportionately received shopping ads from niche sites, while a more mixed population saw shopping ads from sites like Target or Amazon. Especially given the small sample size used here, this is in no way indicative of structured targeting based on this attribute. It seems unlikely, for instance, that lulus.com sought out frequent travelers when choosing their target audience.

[17] Once again, my method does not distinguish between ads served *by* Reddit or Buzzfeed and ads that, for example, linked to articles or promotional pieces hosted on Reddit or Buzzfeed

However, observing these patterns still reveals some interesting co-occurrences. Perhaps shoppers at these sites bias wealthier than those at sites like Amazon or Target, and are therefore more likely to travel. Thus, even if we cannot conclusively

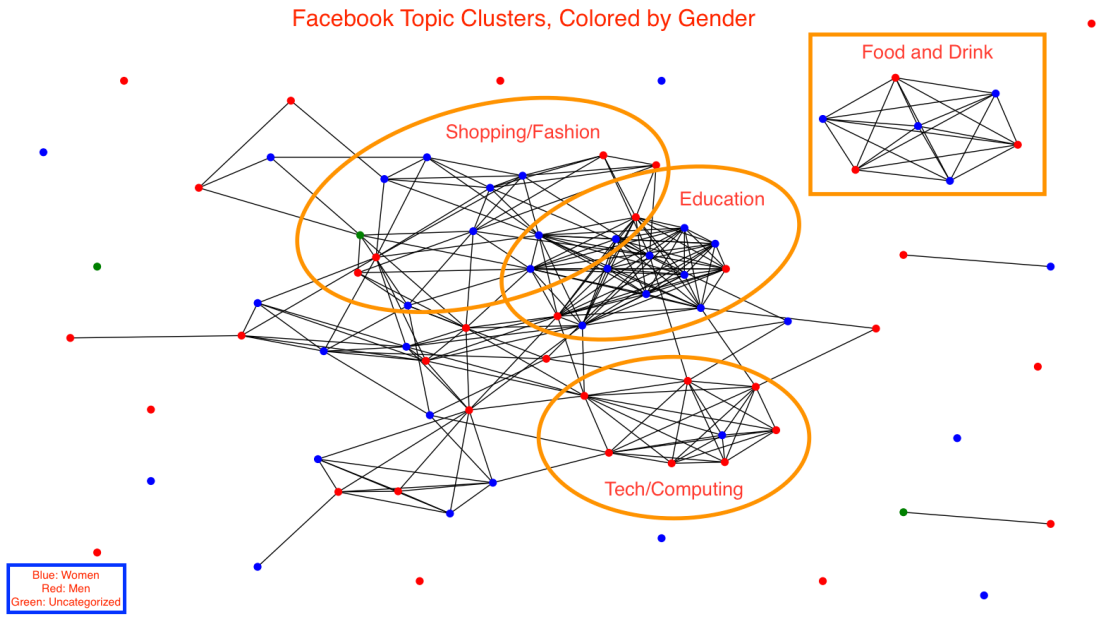


Figure 6.12 : *Demographic clustering graph for Facebook topics and gender*

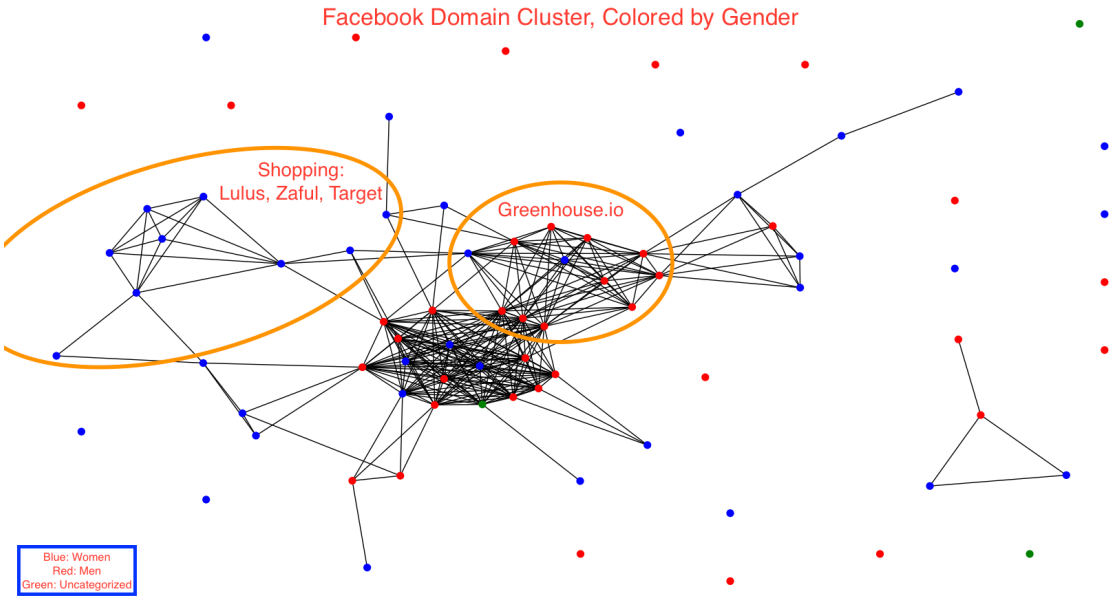


Figure 6.13 : *Demographic clustering graph for Facebook domains and gender*

establish a user's characteristics using this approach, with enough data we could theoretically infer some of their attributes based on the sites that target them most.

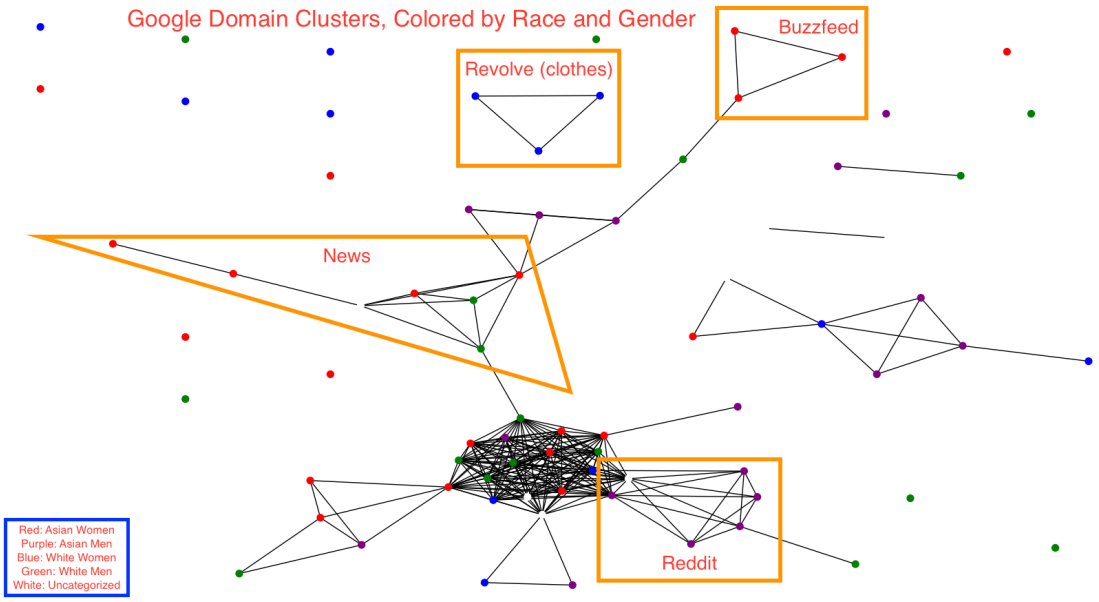


Figure 6.14 : *Demographic clustering graph for Google domains and race/gender*

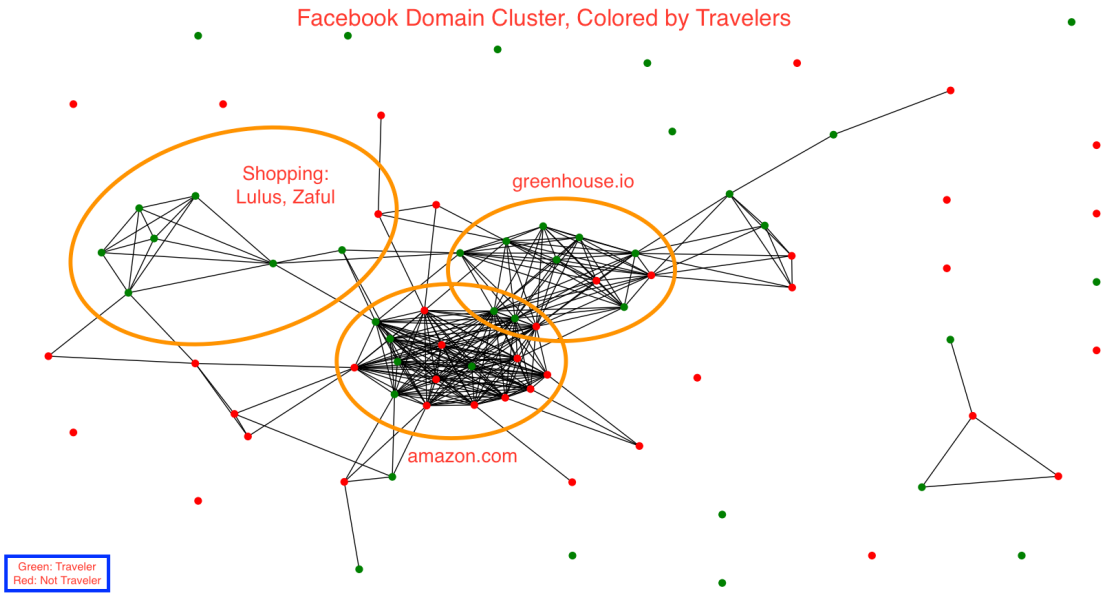


Figure 6.15 : *Attribute clustering for Facebook domains and frequent travelers*

Google		Facebook	
Interest	*Count*	*Interest*	*Count*
Education	60	The New York Times	58
Parenting	59	Business	53
Movies	53	University	48
Politics	48	Spotify	41
Celebrities & Entertainment News	48	Lyft	40

Table 6.3: *Top Google/Facebook interests for participants*

6.2.4 User Interests

The top user interest categories predicted by Google and Facebook for users in my study are displayed in table 6.3. Interestingly, parenting was the second most common Google interest for my largely college-aged study population. The remaining categories generated by Google seem unsurprising; they likely represent the school-related, entertainment, and news sites that college students often visit. On Facebook, three brands make the top-five list for interests, suggesting strong engagement metrics across my sample. Especially given recent news, seeing The New York Times and Lyft on this list isn't all that surprising.

On average, users self-reported **28.9 Google interests** and **82.8 Facebook interests**. The Facebook interest figures are likely to be an underestimate; I asked users to submit the interests from their top 5 interest categories, but many users (myself included) had as many as 15 categories. These differences are indicative of the contrast in each company's interest classification strategy (see figure 6.22); Google classified users into a set of pre-defined buckets, while Facebook used a set of topics and brands to describe users.

On the whole, my interest reclassification attempts were relatively successful. While precision and recall rates on the Google interest set traded off with one another based on the number of top interests categories chosen (see figure 6.16), my techniques had high accuracy rates across the board. On average, of the top 10 interest suggestions for each user, **71% were accurate** (ie. represented in Google's own set); they together captured an average of **26% of a user's Google-defined interest list** (F-Score 38.06). Restricting to the top 5 proposed interests drove **accuracy rates up to 87%** with an average recall of 14% (F-Score 36.40).

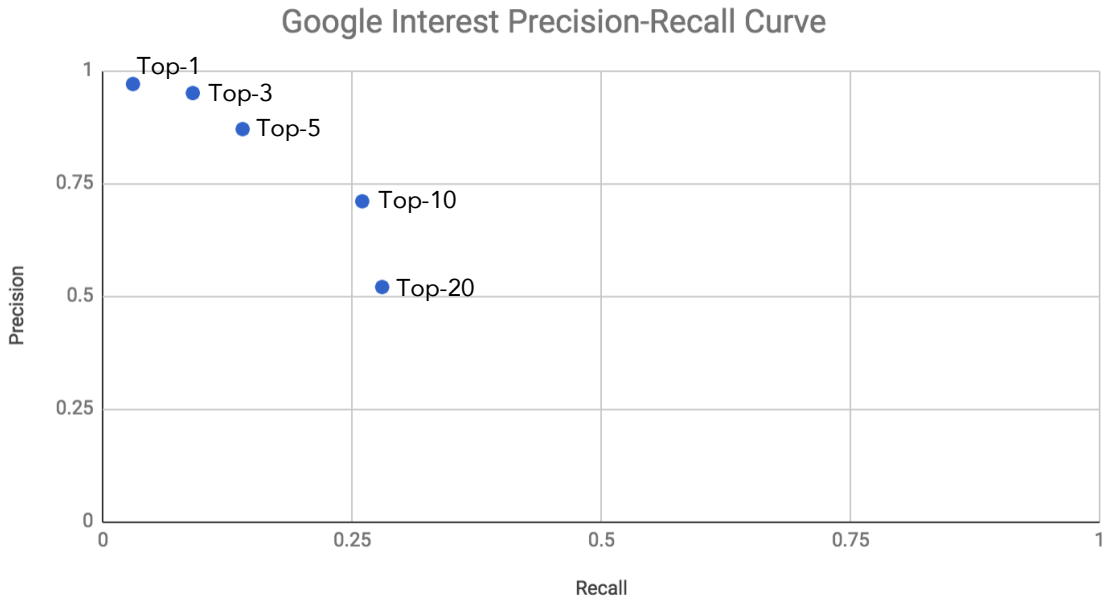


Figure 6.16: *Precision-Recall curve for Google interest re-classification*

To benchmark these results, I used "Betrayed by Your Ads" ([55]), which also attempted to reclassify user interests from advertisements. "Betrayed" differs from my work in a few key ways – most saliently, the authors repeatedly simulate site hits from a set of limited user browsing histories. Since they only access 30 training sites (during profile creation) and between 10-15 test sites (for ad collection), they likely generate limited Google interest profiles, much narrower in

scope than those generated for users with years of browsing activity. Furthermore, though I likely collect far more ads per user on average (since users in my study browse more than 15 sites over the two week collection period), the presence of many potential confounds in my own study should also theoretically complicate my interest reclassification abilities.

The authors of "Betrayed" use three sets of overlap rules to measure reclassification accuracy – commonalities in actual categories, in parent categories, and in root categories. This meant that under the second and third rules, they would count a proposed interest in "Beaches & Islands" as accurate if it shared its parent ("Tourist Destinations") or its root ("Travel") with any interest in the ground-truth set. Given my manually-created correspondence list, neither of these rules perfectly matched my approach; the most reasonable comparison to their work is with their 'parent' rule. Finally, "Betrayed" runs their experiments in one of two scenarios representing different threat vectors they identify – I benchmark my work against their most successful setups.

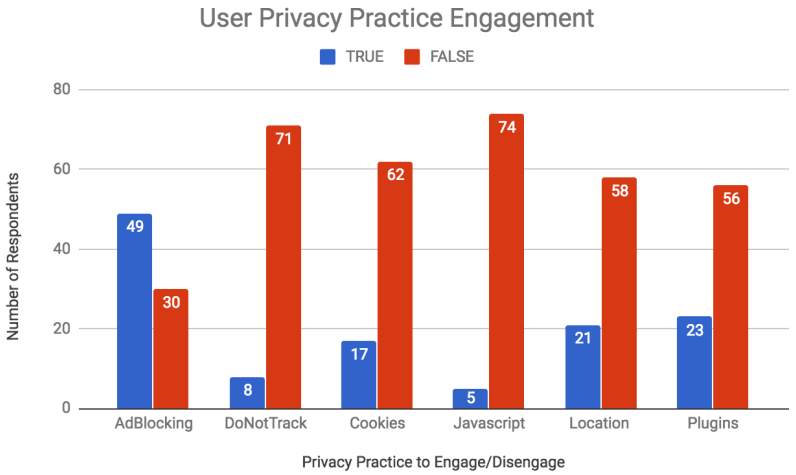
In their "Workplace Scenario" experiments run with 30 training sites and 15 test sites, "Betrayed" achieves precision and recall rates of 45% and 34%, respectively (F-Score 38.59). Though this represents an F-Score marginally higher than my top-10 approach, this is likely driven in part by their higher recall rates which are, as I mentioned, more attainable due to the limited nature of their generated interest sets. Their highest precision rate was 54% ("Hotspot Scenario" with 10 test sites); my top-20 approach yielded similar precision rates with significantly higher recall figures (28% to 20%). Though this is by no means a perfect benchmark, I demonstrate that this technique **achieves similar performance on far more complicated user profiles**, indicating that user interest reclassification is possible even on live users, despite all the complexities therein.

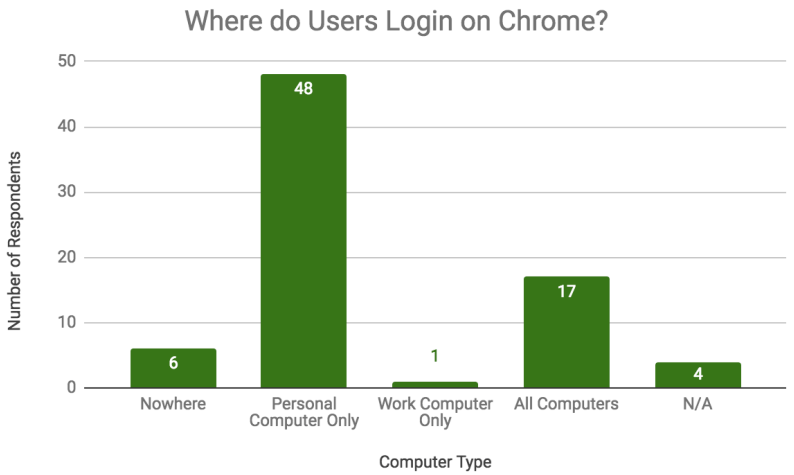
6.2.5 Privacy Outlooks

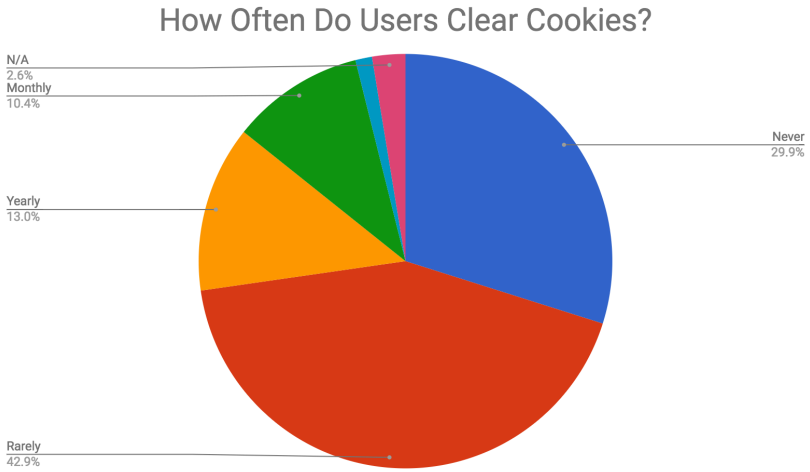
Examining survey responses revealed a lot about user behaviors and outlooks on privacy. Surprisingly (given the assumed technological literacy and interest of my audience – see section overview), **only 62.02% of my users used adblockers** (figure 6.17). Though I do not have more granular data on which adblockers were most popular, I anecdotally found in talking to potential respondents that AdBlock Plus was widely used. This choice is important, as section §3.2 demonstrated; not all adblockers block trackers. Intuitively, we might assume that adblockers in general have some effect on the ability of companies to categorize user interests or tailor ads. To investigate if this was the case, I ran a t-test on the relationship between adblocker use and various measures of ad targeting success (table 6.4).

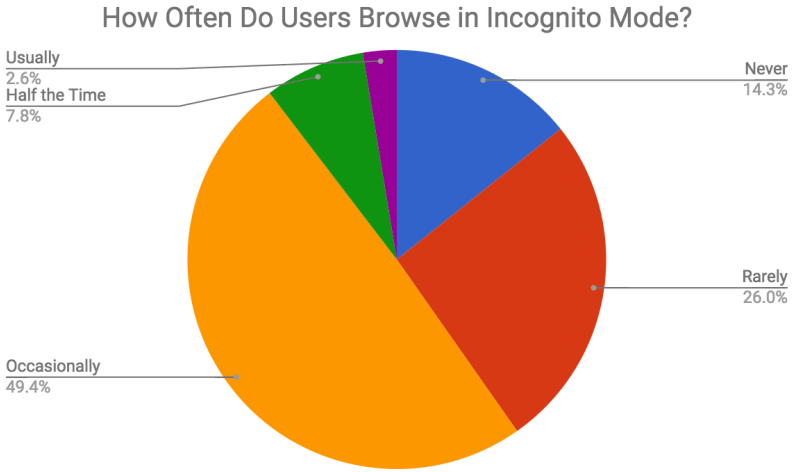
The measures I chose to test included both subjective and objective indicators of quality. This was done to tease out user response effects – users who typically use adblock, for instance, might be more likely to rate ads lower or higher as a class in subjective quality assessments. By including Google's age and gender accuracy, though, I tested whether these two objective indicators of quality were different for adblock users.

Contrary to my working assumption, **I found no statistically significant differences due to adblockers** across the board. For Google measures, in fact, the subset of users with adblock rated their categorizations more highly! This outcome is understandable. Given that some adblockers do not prevent trackers and the fact that Google and Facebook have many other sources of information on users than simply ad engagement (including, for Google, browsing histories on Chrome), it is reasonable that both services can accurately understand users across the board. These results do not conclusively establish that adblockers aren't useful, though. I do not test, for example, the effects of specific blockers or attempt to correct for









Clockwise, from top left: Figures 6.17: *User privacy practice engagement rates;* 6.18: *User login locations on Chrome;* 6.19: *User cookie clearance frequency;* 6.20: *User incognito mode usage frequency*

Table 6.4: *Adblocker effects on perceived interest set quality and ad tailoring frequency*

		Adblock Subset		*No Adblock Subset*		*Sample Std Err*	*No - Yes (t-stat)*
		n	*Mean (Std Err)*	*n*	*Mean (Std Err)*		
General	Ad Tailoring	47	3.532 (0.856)	27	3.556 (0.974)	0.225	0.024 (0.917)
Google	Interest Accuracy	45	3.289 (0.869)	28	3.000 (0.981)	0.226	-0.289 (-1.277)
	Interest Comprehensiveness	45	3.090 (1.080)	28	3.000 (1.150)	0.270	-0.090 (-0.330)
	Age Accuracy	49	0.429 (0.500)	29	0.379 (0.494)	0.116	-0.049 (-0.424)
	Gender Accuracy	49	0.531 (0.504)	29	0.517 (0.509)	0.119	-0.013 (-0.113)
Facebook	Interest Accuracy	41	3.293 (1.055)	26	3.346 (1.018)	0.259	0.053 (0.207)
	Interest Comprehensiveness	41	3.000 (1.095)	26	3.000 (1.095)	0.275	0.000 (0.000)
	Category Accuracy	39	2.939 (1.842)	27	3.483 (1.379)	0.397	0.544 (1.371)

* significant at $p < 0.05$; ** significant at $p < 0.01$; *** significant at $p < 0.001$; **** significant at $p < 0.0001$

misplaced causation – it could be the case that users engaged adblockers *because* they were unnerved by shockingly accurate ads. More rigorous studies on the matter would need to collect more granular information in a controlled setting. Furthermore, ad blockers might be effective at blocking ads even if they aren't effective in preventing tracking or protecting privacy.

Other survey responses on privacy practices were illuminating as well. As could be expected, far more users block plugins or location tracking by default than block cookies or Javascript. Blocking Javascript and cookies would interfere with the basic functioning of many sites on the web. Far more concerning is the fact that **almost 75% of these users never or rarely clear their cookies**. Even if techniques like fingerprinting and syncing mean that clearing cookies isn't a perfect way of fighting tracking, it is one simple tool at a user's disposal that isn't too difficult to engage. Meanwhile, less than 10% of users noted that they browsed in incognito mode more often than not. Again, while incognito mode doesn't stop third parties from learning about users, it might be one way of reducing flows to Google. Figures 6.19 and 6.20 display more granular representations of this information.

My study sample typically logs on to Chrome only on their personal computers, reducing the likelihood of partial ad observation at work mentioned earlier (figure 6.18). That said, these responses should be taken with a grain of salt; as noted earlier, most of my respondents are in college and therefore may not have work computers or may not trust the computers they receive only for internships.

I also asked participants to indicate whether they thought the ads they saw on the internet were successfully tailored to their interests and to what degree they wanted them to be (see figure 6.21). This question pair yielded two interesting conclusions – users broadly perceived ads as accurately being tailored to their interests, but wanted a lower amount of interest-based targeting than they experienced in the status quo.

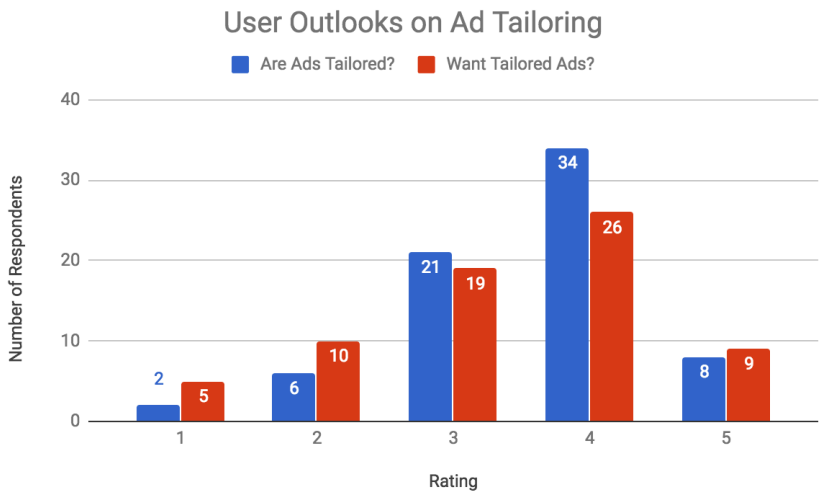


Figure 6.21: *User outlooks on ad tailoring*

6.2.6 Site Differences

Finally, I turned to investigating the differences between ad agents. I first examined the comparative quality of Google's and Facebook's interest classifications. The two sites take very different approaches to interest categorization: Google uses general topics from a pre-defined list, while Facebook has assortments of categories, ideas, and products (see figure 6.22). As such, we might expect that users' ratings of their respective accuracy and comprehensiveness would differ (especially since users saw both lists and could choose their ratings after directly comparing them). To see if this was the case, I ran a t-test to compare user reported accuracy and comprehensiveness assessments for Google and Facebook interest lists (see table 6.5).

Surprisingly, there were **no statistically significant differences in user ratings** of the two lists. Facebook was judged as more accurate while Google was more comprehensive, but both differences were marginal. Naturally we cannot conclude from these data that Facebook and Google are equally good at understanding users – each may have particular kinds of users they're uniquely well suited at reading,

for example – but they do indicate that we should be cautious about assuming that one or the other is better.

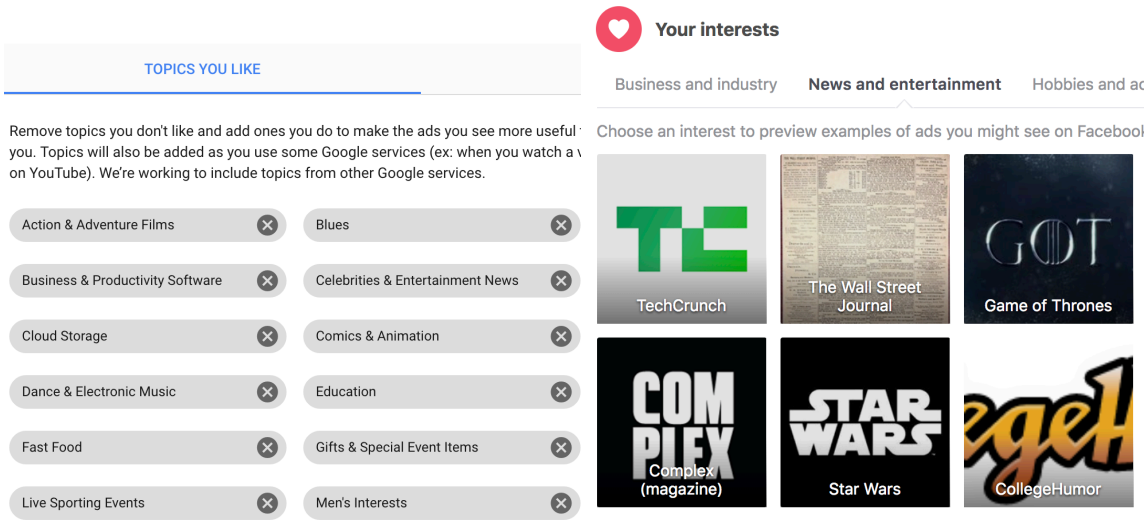


Figure 6.22: *My interests, as identified by Google (left) and Facebook (right)*

With these results in mind, I then examined user faith in the two services. I first plotted user trust in Facebook, Google, and third party advertisers (figure 6.23). Immediately, a few conclusions became apparent. Most users indicated the lowest possible trust rating for third parties and none trusted them with a rating above 3 out of 5. This suggested that the major duopoly had somehow managed to distinguish themselves from the general distrust of the advertising ecosystem. Google and Facebook trust indicators were more measured, with users in each of the five buckets. That said, their respective trend lines indicated that Google was slightly more trusted on net, with a distribution more even than Facebook's. In order to see how substantial these differences were, I ran t-tests on trust averages for each entity (table 6.6).

Table 6.5: *Facebook vs. Google – profile quality measures*

	Facebook		*Google*		*Sample Std Err*	*FB – Google (t-stat)*
	n	*Mean (Std Err)*	*n*	*Mean (Std Err)*		
Accuracy Ratings	70	3.300 (1.026)	73	3.178 (0.918)	0.163	0.122 (0.748)
Comprehensiveness Ratings	70	3.000 (1.077)	73	3.055 (1.104)	0.182	-0.055 (-0.300)

* significant at $p < 0.05$; ** significant at $p < 0.01$; *** significant at $p < 0.001$; **** significant at $p < 0.0001$

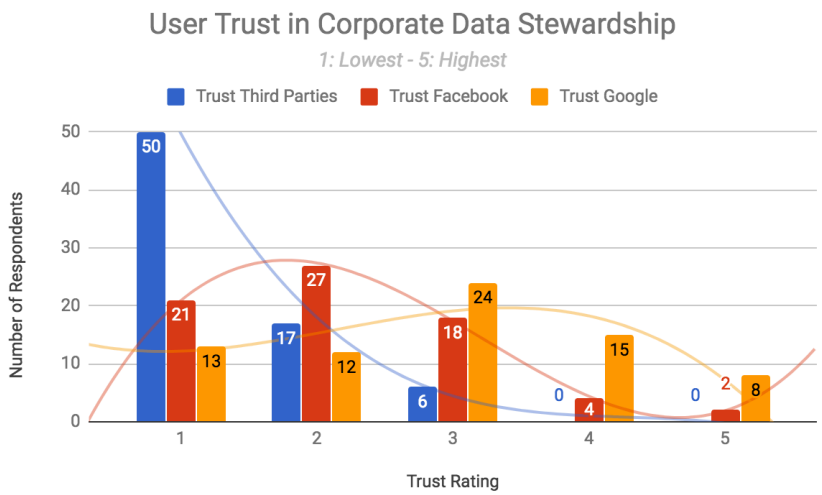


Figure 6.23 : *User trust in Google, Facebook, and third party advertisers*

This time, all differences were significant. **Google was trusted substantially more than other advertisers, and Facebook more than third parties**. This wasn't surprising (especially given recent news), but is still a meaningful indicator of how factors external to categorization ability can affect perceptions of advertisers. To confirm these results on my set, I plotted whether my respondents differed in their comfort with Facebook, Google, or third parties handling their interests (see figure 6.24).

Perhaps due to the middling overall indications of interest list quality (accuracy and comprehensiveness ratings averaged between 3 and 3.3 for both services) and since interests likely seemed innocuous enough, the majority of respondents (59.5%) indicated they felt no discomfort at sharing their interests with Google/Facebook or other advertisers. This does not give these companies a clean sheet, however. These firms were still generally mistrusted (all three trust averages were below 3 out of 5) and these findings do not take into account user awareness of how abstract interest categorizations could be used to create specific pictures of user needs. I found that 31.1% of respondents trusted Google and Facebook with interests they'd feel uncomfortable sharing with third-party

Table 6.6: *Facebook vs. Google vs. third party advertisers – trust measures*

	Facebook		*Google*		*Third Parties*		*Sample Std Err*	*First - Second (t-stat)*
	n	*Mean (Std Err)*	*n*	*Mean (Std Err)*	*n*	*Mean (Std Err)*		
Google vs. Facebook	74	2.176 (1.012)	74	2.946 (1.259)			0.188	0.770**** (4.102)
Facebook vs. Third Parties	74	2.176 (1.012)			74	1.405 (0.639)	0.139	0.770**** (5.537)
Google vs. Third Parties			74	2.946 (1.259)	74	1.405 (0.639)	0.164	1.541**** (9.383)

* significant at p $<$ 0.05; ** significant at p $<$ 0.01; *** significant at p $<$ 0.001; **** significant at p $<$ 0.0001

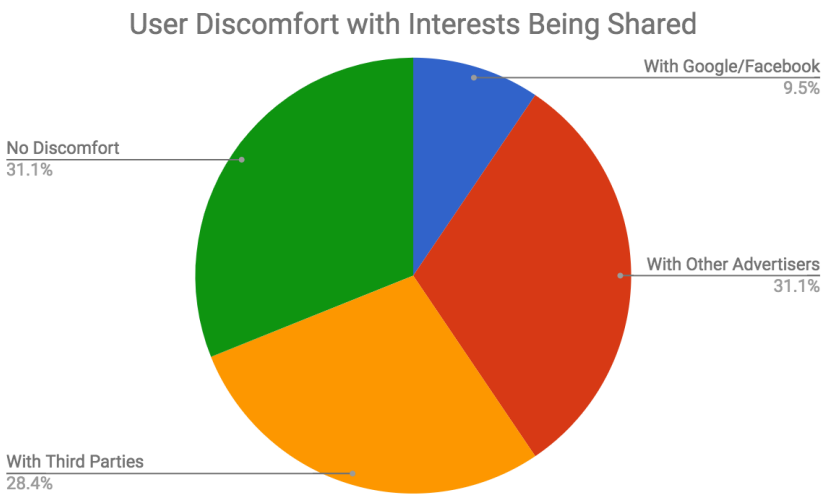


Figure 6.24: *User discomfort with interests being shared with other parties*

advertisers, once again indicating substantial differences between the two sets. I also found that **9.5% of respondents were uncomfortable with Google/Facebook having the interest classifications they'd already compiled**.

Finally, I investigated how different user metrics were correlated with one another (see table 6.7). I began by analyzing how metrics on interest/category accuracy or comprehensiveness affected user trust in Facebook or Google. I found that Facebook's category accuracy and Google's interest comprehensiveness, respectively, had the greatest correlation with trust in each service. This does not imply a causal link – in fact, it may well be that the users who trust these services the most end up willingly providing information to them that enables better classification, or that users who saw better classifications were as a result more likely to trust the two.

More interestingly, I investigated how each indicator was correlated to whether users found that advertisements on the internet were tailored to them. Here, what was surprising were not the individual figures (though the fact that Facebook category accuracy on its own was strongly correlated with tailoring assessments was unexpected), but rather the differences between these correlations

for Google and Facebook. While **all three Facebook metrics were well correlated with tailoring assessments of online ads**, **none of the Google metrics were**. One particularly compelling explanation for this phenomenon might suggest that users are comparatively more aware of Facebook advertisements when thinking about targeted ads on the web. Anecdotally, I've found that this is the case for many individuals in my community. If true, this would have interesting ramifications for Facebook, especially given the trust conclusions discussed earlier.

6.3 Case Studies

In this section, I explore two case studies, analyzing on a deeper level what this dataset can reveal for individual users and assessing one possible policy implication from my survey results.

6.3.1 User Information Studies

Thus far, in my discussion of sensitive ad data leakage, I aggregated study-wide statistics that validated risks across all participants. To further contextualize this issue, however, I now turn to the two users with the highest sensitive ad content hit rates on Google and Facebook, respectively, in order to demonstrate just how revealing these data can be.

The Facebook user who received the most sensitive ad hits saw disproportionately high numbers of ads from religious organizations – 257 ads shown to this participant were about religion or spirituality. What's more, manually examining some of these domains reveals that the vast majority of these ads were for Jewish cultural organizations. Separately, this user received ads from conservative groups like the Network of Enlightened Women, indicating a likely political stance as well. Though I do not have a ground truth metric with which to

Table 6.7: *Facebook, Google, and privacy outlooks – correlation matrix*

	Facebook_Trust *On a 1-5 scale, "I trust Facebook as a steward of my personal data"*	*Google_Trust* *On a 1-5 scale, "I trust Google as a steward of my personal data"*	*Is_Tailored* *On a 1-5 scale, "...I find that the ads I see are tailored to my interests"*
FB_Acc *On a 1-5 scale, "How accurate [are the FB-generated] interests?"*	0.078 (0.006)		0.280 (0.078)
FB_Comp *On a 1-5 scale, "How comprehensive [are the FB- generated] interests?"*	0.134 (0.018)		0.384 (0.148)
FB_Cat *On a 1-5 scale, "How accurate [are the FB-generated] categories?"*	0.290 (0.084)		0.370 (0.137)
FB_Avg *Average of the above metrics (used to indicate quality)*	0.238 (0.056)		0.489 (0.239)
Goog_Acc *On a 1-5 scale, "How accurate [are the Google-generated] interests?"*		0.172 (0.030)	-0.104 (0.011)
Goog_Comp *On a 1-5 scale, "How comprehensive [are the Google-generated] interests?"*		0.213 (0.046)	-0.018 (0.000)
Goog_Avg *Average of the above metrics (used to indicate quality)*		0.217 (0.047)	-0.063 (0.004)

All table values represent Pearson correlation coefficients. R^2 *in parentheses*

judge these inferences, observing the Facebook interest list for this user revealed categories that were closely related to the ads shown, including Israel, the Republican Party, and motherhood. The fact that these ads so clearly painted a stereotypical picture of this user that was then corroborated by Facebook's interest list is worrying, insofar as it illustrates just how revealing ads can be if observed manually.

On Google, the user with the highest number of sensitive ad hits was largely targeted by masculine health and fitness sites. Bodybuilding, general health, men's health, and alternative medicine related ads all appeared frequently for this user's profile. Separately, this participant also saw five ads related to weapons and five for sites with adult content. Together these ads paint a similarly concerning picture. Though this user's Google interests did not include categories related to the above ads, their Facebook list revealed interests in first person shooter games, 'perfection', 'health and wellness', and 'adult'. Once again, the stereotypical picture suggested by this user's ad hits could be used to paint a somewhat accurate portrayal of their interests.

6.3.2 Browsing Behavior Studies

On the whole, the survey responses I discuss in sections §6.2.5 and §6.2.6 suggest that users aren't fully content with the state of the online advertising ecosystem in the status quo. Trust metrics average below 3 out of 5 for Google, Facebook, and third party advertisers alike; most users seem to want ad tailoring less than they currently see it.

Furthermore, 36% of users expressed surprise at seeing their personal ad profiles on Google or Facebook – either at its existence or at its accuracy (see figure 6.25). Taken together, this might, at first glance, suggest that these users might be ready to change how they browse the web or engage with these platforms.

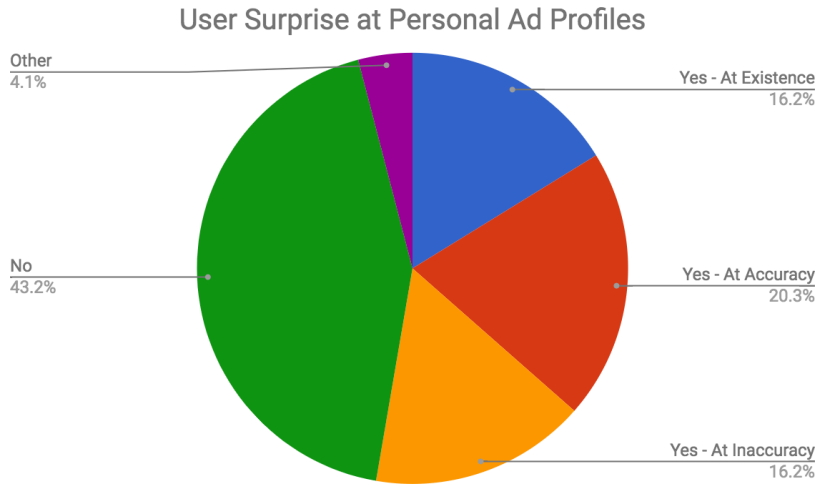


Figure 6.25: *User surprise at seeing Facebook/Google ad profiles*

As figure 6.26 indicates, however, this is not the case. Users' reluctance to change their browsing behaviors illuminates an important piece of policy discussions regarding online privacy protection. User buy-in is key; future solutions must convince regular consumers that they can easily modify their behaviors to realize marked improvements in privacy.

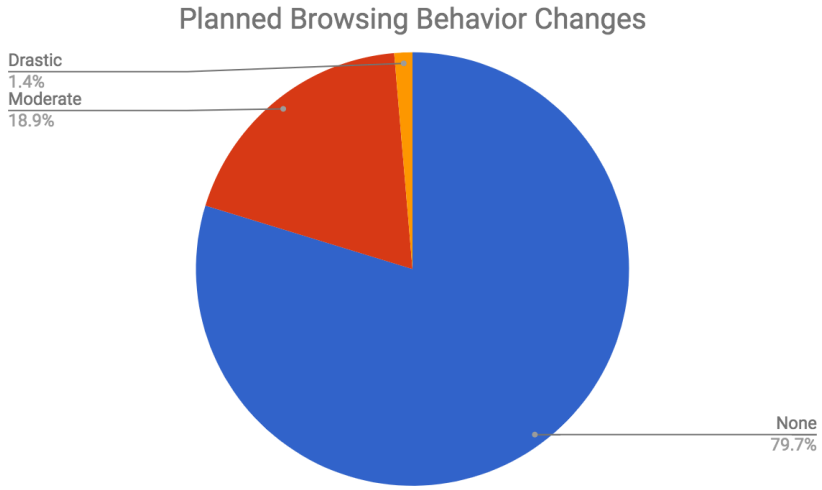


Figure 6.26: *Users' planned browsing behavior changes*

Table 6.8: *Summary of select findings*

Key Question	**Technique/Topic**	**Select Finding(s)**
Personal Identifiers	PII flag counts	1.44% of ads leaked current location, 0.33% leaked names
	Manual inspection	Multiple ads found with de-anonymization potential
Sensitive Sites	HTTP link/content counts	At least one HTTP link on 29% of Facebook, 13% of Google ads
	Sensitive ad content	About 6% of ads sensitive; proportion varied by user
	Malicious site counts	No malicious site links found
	Manual inspection	Instances of locally sensitive ads found
Demographic Information	Gender classification module	47 accurate gender predictions made (23 for users w/o accurate Google predictions)
	Gender/race graphing	Ad makeup differs by gender and race in implicitly sexist ways
	Facebook attribute graphing	Co-occurrence of attributes like traveling and niche site shopping
User Interests	Interest classification module	71% precision/26% recall rates for Google interests
		Manually-observed interest matches for Facebook interest set
Privacy Outlooks	Adblocker use	62% of participants used adblockers, no difference in interest accuracy, comprehensiveness, or ad tailoring ratings
	Cookie clearing	75% of participants never or rarely clear their cookies
Site Differences	Profile quality tests	No significant differences in user profile ratings
	Agent trust tests	Users trusted Google more than Facebook, both over 3rd parties
	Tailoring correlations	Facebook ratings correlated with tailoring perception
	Planned changes	No planned changes for most users

CHAPTER 7

CONCLUSION

On May 7th, almost a month after Facebook's interrogation in the Senate, an editorial in the Wall Street Journal declared that privacy was dead. "Short of living in a remote hut... there is no longer any way that you... can prevent marketers, governments or malicious actors from gathering and using comprehensive, personally identifying information about you." [92] Whether we ought to be so pessimistic about our privacy is still an open question. The efforts of those in university research labs and legislative chambers alike may yet generate genuine safeguards for our data. What is undeniable, though, is the importance of advertising as a piece in the puzzle of online privacy.

In this thesis, I set out to understand what we could learn about advertisers and data flows from investigating targeted ads. The online advertising ecosystem is complex, and though it is a well-studied space overall, past research has left

many holes in our understanding of real-world targeting. In particular, the striking lack of live-user studies in the existing literature meant that we did not have an up-to-date picture of how data leakages occur from complex, detailed user profiles or on Facebook. To fill these gaps, I developed a novel approach that combined both orchestrated and live data in what is, to my knowledge, the largest live-user study of targeted ads to date.

This method is not without its limitations. As I warn, my results do not necessarily hold for samples outside of my own set. In many ways, my study population is non-representative of both Princeton University students and Google/Facebook users more generally. Furthermore, although I demonstrated the ability to reconstruct pieces of a user's profile from the ads targeted to them, I did not make any claims about whether those ads were targeted based on these attributes. My methods proved effective, to a degree, on my relatively homogenous study participants; it remains to be seen if they can successfully be generalized to more diverse samples.

With that said, the data I collected suggested some disturbing conclusions. I observed both personal identifier leakage and sensitive topic references in ads that often included HTTP-served (insecure) content or links. These findings validated both threat vectors described in Chapter 4, indicating that unauthorized agents could theoretically leverage advertisements in order to harm unsuspecting users. By running more complex scripts on the ads I collected, I was able to reclassify many participants' demographic characteristics and interests. Graphing ad targeting behavior for various user attributes also demonstrated that the makeup of ads a user sees can differ sharply based on seemingly irrelevant features and often in implicitly sexist ways. Finally, analyzing user outlooks on privacy and advertising indicated that the users in my sample differed sharply on everything from the privacy measures they adopted to their degree of comfort with advertisers.

In Chapter 4, I outlined a three-pronged ethical standard that I argued we should apply to Google and Facebook. Assessed on this metric, these findings indicate that the duopoly may not be conforming to legal regulations, industry norms, and customer expectations on data use transparency. At the very least, this research could illuminate a starting point for future work in dynamically reverse engineering data usage for ads in real time.

There is much we still do not know about online advertising. Live user studies could add a valuable layer to our understanding of how location or site retargeting affects advertising choices. More rigorous studies of privacy-protective measures engaged by users could help identify the techniques that work best. Finally, although doing so is methodologically harder, observing how third party ad agents collect and use information could be transformative for our assessment of online privacy risks. As ad agents continue to exercise a larger and larger influence on the internet, we must continue to refine our understanding of the online advertising ecosystem, its hidden nuances, and the threat it poses to privacy.

References:

[1] "A timeline of Facebook's privacy issues — and its responses," *NBC News*. [Online]. Available: https://www.nbcnews.com/tech/social-media/timeline-facebook-s-privacy-issues-its-responses-n859651. [Accessed: 02-May-2018].

[2] T. courtesy of B. Government, "Transcript of Mark Zuckerberg's Senate hearing," *Washington Post*, 10-Apr-2018.

[3] D. Thompson, "Facebook and Google Own the Future of Advertising—in 2 Charts," *The Atlantic*, 25-Mar-2014.

[4] P. Kafka, "2017 was the year digital ad spending finally beat TV," *Recode*, 04-Dec-2017. [Online]. Available: https://www.recode.net/2017/12/4/16733460/2017-digital-ad-spend-advertising-beat-tv. [Accessed: 28-Mar-2018].

[5] N. Singer, "Your Online Attention, Bought in an Instant by Advertisers," *The New York Times*, 17-Nov-2012.

[6] "As Brands Turn to Digital Advertising to Reach the Right Audience, Focus on Validation Is Increasing." [Online]. Available: https://www.forbes.com/sites/forbespr/2015/05/05/as-brands-turn-to-digital-advertising-to-reach-the-right-audience-focus-on-validation-is-increasing/#79c3faec272c. [Accessed: 20-Apr-2018].

[7] "What Is An Ad Network? - MarTech Landscape," *MarTech Today*, 30-Dec-2015. [Online]. Available: https://martechtoday.com/martech-landscape-what-is-an-ad-network-157618. [Accessed: 28-Mar-2018].

[8] "What Is an Ad Network and How Does It Work? - Clearcode Blog," *Clearcode - Enterprise-grade Software Development*, 07-Mar-2018. [Online]. Available: https://clearcode.cc/blog/what-is-an-ad-network-and-how-does-it-work/. [Accessed: 25-Mar-2018].

[9] "What Is An Ad Exchange? - MarTech Landscape," *MarTech Today*, 01-Feb-2016. [Online]. Available: https://martechtoday.com/martech-landscape-what-is-an-ad-exchange-161947. [Accessed: 28-Mar-2018].

[10] "What is an Ad Server and How Does It Work? - Clearcode Blog," *Clearcode - Enterprise-grade Software Development*, 07-Mar-2018. [Online]. Available: https://clearcode.cc/blog/what-is-an-ad-server/. [Accessed: 25-Mar-2018].

[11] J. Davies, "Know your cookies: A guide to internet ad trackers," *Digiday*, 01-Nov-2017. .

[12] J. BohannonMar. 11, 2013, and 3:00 Pm, "Facebook Preferences Predict Personality Traits," *Science | AAAS*, 11-Mar-2013. [Online]. Available: http://www.sciencemag.org/news/2013/03/facebook-preferences-predict-personality-traits. [Accessed: 07-Feb-2018].

[13] J. BohannonJan. 12, 2015, and 3:30 Pm, "Your computer knows you better than your friends do," *Science | AAAS*, 12-Jan-2015. [Online]. Available: http://www.sciencemag.org/news/2015/01/your-computer-knows-you-better-your-friends-do. [Accessed: 07-Feb-2018].

[14] J. C. Wong, "'It might work too well': the dark art of political advertising online," *the Guardian*, 19-Mar-2018. [Online]. Available: http://www.theguardian.com/technology/2018/mar/19/facebook-political-ads-social-media-history-online-democracy. [Accessed: 20-Apr-2018].

[15] "What is Cookie Syncing and How Does it Work? - Clearcode Blog." [Online]. Available: https://clearcode.cc/blog/cookie-syncing/. [Accessed: 25-Mar-2018].

[16] "The hidden perils of cookie syncing." [Online]. Available: https://freedom-to-tinker.com/2014/08/07/the-hidden-perils-of-cookie-syncing/. [Accessed: 27-Mar-2018].

[17] "Deterministic and Probabilistic Matching: How Do They Work? - Clearcode Blog," *Clearcode - Enterprise-grade Software Development*, 15-Dec-2016. [Online]. Available: https://clearcode.cc/blog/deterministic-probabilistic-matching/. [Accessed: 25-Mar-2018].

[18] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild," 2014, pp. 674–689.

[19] S. Englehardt and A. Narayanan, "Online Tracking: A 1-million-site Measurement and Analysis," 2016, pp. 1388–1401.

[20] "Study: Use of ad blockers increases 16% in US," *Marketing Dive*. [Online]. Available: https://www.marketingdive.com/news/study-use-of-ad-blockers-increases-16-in-us/507922/. [Accessed: 28-Mar-2018].

[21] "The New Chrome and Safari Will Reshape the Web," *WIRED*. [Online]. Available: https://www.wired.com/2017/06/new-chrome-safari-will-reshape-web/. [Accessed: 28-Mar-2018].

[22] "Yelp offers new ad customization to advertisers," *Search Engine Land*, 02-Mar-2018. [Online]. Available: https://searchengineland.com/yelp-offers-new-ad-customization-advertisers-293317. [Accessed: 27-Mar-2018].

[23] "Advertising Personalization and Landing Pages: The Next Wave in Digital Marketing," 28-Oct-2016. .

[24] "The Truth About Online Privacy: How Your Data is Collected, Shared, and Sold - Clearcode Blog," *Clearcode - Enterprise-grade Software Development*, 07-Sep-2015. [Online]. Available: https://clearcode.cc/blog/online-privacy-user-data/. [Accessed: 25-Mar-2018].

[25] A. Rodgers, "Display retargeting tags are present on more than half of top websites," *Econsultancy*, 31-Mar-2014. [Online]. Available:

https://www.econsultancy.com/blog/64509-display-retargeting-tags-are-present-on-more-than-half-of-top-websites. [Accessed: 27-Mar-2018].
[26] "About ad customizers - AdWords Help." [Online]. Available: https://support.google.com/adwords/answer/6072565?hl=en. [Accessed: 27-Mar-2018].
[27] "Data collection and use - Advertising Policies Help." [Online]. Available: https://support.google.com/adwordspolicy/answer/6020956?hl=en. [Accessed: 07-Feb-2018].
[28] L. Bright and T. Daugherty, "Does customization impact advertising effectiveness? An exploratory study of consumer perceptions of advertising in customized online environments," *J. Mark. Commun.*, vol. 18, pp. 19–37, Feb. 2012.
[29] S. Guha, B. Cheng, and P. Francis, "Challenges in measuring online advertising systems," 2010, p. 81.
[30] A. Korolova, "Privacy Violations Using Microtargeted Ads: A Case Study," p. 24.
[31] M. Conti, V. Cozza, M. Petrocchi, and A. Spognardi, "TRAP: Using Targeted ads to unveil Google personal profiles," 2015, pp. 1–6.
[32] X. Xing, W. Meng, D. Doozan, A. C. Snoeren, N. Feamster, and W. Lee, "Take This Personally: Pollution Attacks on Personalized Services," p. 16.
[33] C. Castelluccia, E. De Cristofaro, and D. Perito, "Private Information Disclosure from Web Searches," in *Privacy Enhancing Technologies*, vol. 6205, M. J. Atallah and N. J. Hopper, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 38–55.
[34] S. Englehardt, J. Han, and A. Narayanan, "I never signed up for this! Privacy implications of email tracking," *Proc. Priv. Enhancing Technol.*, vol. 2018, no. 1, pp. 109–126, Jan. 2018.
[35] J. A. Calandrino, A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov, "'You Might Also Like:' Privacy Risks of Collaborative Filtering," 2011, pp. 231–246.
[36] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," 2009, p. 7.
[37] C. Eubank, M. Melara, D. Perez-Botero, and A. Narayanan, "Shining the Floodlights on Mobile Web Tracking — A Privacy Survey," p. 9.
[38] S. Englehardt, C. Eubank, P. Zimmerman, D. Reisman, and A. Narayanan, "Web Privacy Measurement: Scientific principles, engineering platform, and new results," p. 17.
[39] F. Roesner, T. Kohno, and D. Wetherall, "Detecting and Defending Against Third-Party Tracking on the Web," p. 14.
[40] S. Schelter and J. Kunegis, "On the Ubiquity of Web Tracking: Insights from a Billion-Page Web Crawl," *J. Web Sci.*, vol. 4, no. 4, pp. 53–66, Mar. 2018.

[41] S. Macbeth, "Tracking the Trackers: Analysing the global tracking landscape with GhostRank," p. 13.
[42] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting," 2013, pp. 541–555.
[43] G. Acar *et al.*, "FPDetective: dusting the web for fingerprinters," 2013, pp. 1129–1140.
[44] M. H. Mughees, Z. Qian, and Z. Shafiq, "Detecting Anti Ad-blockers in the Wild," *Proc. Priv. Enhancing Technol.*, vol. 2017, no. 3, Jan. 2017.
[45] B. Krishnamurthy, K. Naryshkin, and C. E. Wills, "Privacy leakage vs. Protection measures: the growing disconnect," p. 10.
[46] "Tracking the Trackers: Where Everybody Knows Your Username." [Online]. Available: /blog/2011/10/tracking-trackers-where-everybody-knows-your-username. [Accessed: 19-Apr-2018].
[47] S. Englehardt *et al.*, "Cookies That Give You Away: The Surveillance Implications of Web Tracking," 2015, pp. 289–299.
[48] L. Olejnik, M.-D. Tran, and C. Castelluccia, "Selling off Privacy at Auction," 2014.
[49] "The Moment of Truth Has Come for Digital Advertising's Transparency Problem." [Online]. Available: http://www.adweek.com/digital/the-moment-of-truth-has-come-for-digital-advertisings-transparency-problem/. [Accessed: 20-Apr-2018].
[50] A. Odlyzko, "Privacy, Economics, and Price Discrimination on the Internet," p. 16.
[51] D. Mattioli, "On Orbitz, Mac Users Steered to Pricier Hotels," *Wall Street Journal*, 23-Aug-2012.
[52] L. Sweeney, "Discrimination in Online Ad Delivery," p. 36.
[53] J. van 't Riet *et al.*, "Investigating the Effects of Location-Based Advertising in the Supermarket: Does Goal Congruence Trump Location Congruence?," *J. Interact. Advert.*, vol. 16, no. 1, pp. 31–43, Jan. 2016.
[54] A. Datta, M. C. Tschantz, and A. Datta, "Automated Experiments on Ad Privacy Settings," *Proc. Priv. Enhancing Technol.*, vol. 2015, no. 1, Jan. 2015.
[55] C. Castelluccia, M.-A. Kaafar, and M.-D. Tran, "Betrayed by Your Ads!," in *Privacy Enhancing Technologies*, vol. 7384, S. Fischer-Hübner and M. Wright, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–17.
[56] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: harvesting and analyzing online display ads," 2014, pp. 597–608.
[57] C. E. Wills and C. Tatar, "Understanding What They Do with What They Know (Short Paper)," p. 6.

[58] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan, "AdReveal: improving transparency into online targeted advertising," 2013, pp. 1–7.
[59] M. C. Tschantz, A. Datta, A. Datta, and J. M. Wing, "A Methodology for Information Flow Experiments," 2015, pp. 554–568.
[60] M. Lecuyer *et al.*, "XRay: Enhancing the Web's Transparency with Differential Correlation," p. 16.
[61] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas, "Adnostic: Privacy Preserving Targeted Advertising∗," p. 23.
[62] S. Guha, B. Cheng, and P. Francis, "Privad: Practical Privacy in Online Advertising," p. 14.
[63] G. Storey, D. Reisman, A. Narayanan, and J. Mayer, "The Future of Ad Blocking: An Analytical Framework and New Techniques," p. 17.
[64] P. Laperdrix, B. Baudry, and V. Mishra, "FPRandom: Randomizing Core Browser Objects to Break Advanced Device Fingerprinting Techniques," in *Engineering Secure Software and Systems*, vol. 10379, E. Bodden, M. Payer, and E. Athanasopoulos, Eds. Cham: Springer International Publishing, 2017, pp. 97–114.
[65] R. Balebako, P. G. Leon, R. Shay, B. Ur, Y. Wang, and L. F. Cranor, "Measuring the Effectiveness of Privacy Tools for Limiting Behavioral Advertising," p. 10.
[66] "Tracking the Trackers: Self-Help Tools." [Online]. Available: /blog/2011/09/tracking-trackers-self-help-tools. [Accessed: 19-Apr-2018].
[67] J. Mazel, R. Garnier, and K. Fukuda, "A comparison of web privacy protection techniques," p. 15.
[68] G. Merzdovnik *et al.*, "Block Me If You Can: A Large-Scale Study of Tracker-Blocking Tools," 2017, pp. 319–333.
[69] T. Bujlow, V. Carela-Espanol, B.-R. Lee, and P. Barlet-Ros, "A Survey on Web Tracking: Mechanisms, Implications, and Defenses," *Proc. IEEE*, vol. 105, no. 8, pp. 1476–1510, Aug. 2017.
[70] G. Ip, "The Antitrust Case Against Facebook, Google and Amazon," *Wall Street Journal*, 16-Jan-2018.
[71] Monumetric, "SSL Certificates & Digital Ads," *Monumetric*, 05-Aug-2016. [Online]. Available: https://blog.monumetric.com/ssl-certificates-digital-ads-dont-play-nice-here-s-why-134ce94a86ef. [Accessed: 03-May-2018].
[72] "Mediavine SSL Ads: Should You Go Secure?," *Mediavine*, 29-Nov-2016. .
[73] "Yes, Your Employer Knows Exactly What You're Doing Online," *Money*. [Online]. Available: http://time.com/money/4890303/employer-browsing-data/. [Accessed: 03-May-2018].
[74] O. Solon, "US border agents are doing 'digital strip searches'. Here's how to protect yourself," *The Guardian*, 31-Mar-2017.

[75] R. Epstein, "Privacy and the Third Hand: Lessons from the Common Law of Reasonable Expectations," *Berkeley Technol. Law J.*, vol. 24, no. 3, p. 1199, Jun. 2009.
[76] "What Is the," *Findlaw*. [Online]. Available: http://injury.findlaw.com/torts-and-personal-injuries/what-is-the--reasonable-expectation-of-privacy--.html. [Accessed: 08-Feb-2018].
[77] "Facebook Data Use Policy." [Online]. Available: https://www.facebook.com/full_data_use_policy. [Accessed: 07-Feb-2018].
[78] "Privacy Policy – Privacy & Terms – Google." [Online]. Available: https://www.google.com/policies/privacy/. [Accessed: 07-Feb-2018].
[79] "Key Changes with the General Data Protection Regulation," *EU GDPR Portal*. [Online]. Available: http://eugdpr.org/key-changes.html. [Accessed: 07-Feb-2018].
[80] "Data protection in the United States: overview | Practical Law." [Online]. Available: https://content.next.westlaw.com/6-502-0467?transitionType=Default&firstPage=true&bhcp=1&contextData=(sc.Default). [Accessed: 07-Feb-2018].
[81] "Why Equifax Executives Will Get Away With the Worst Data Breach in History," *Fortune*. [Online]. Available: http://fortune.com/2017/09/16/equifax-legal/. [Accessed: 08-Feb-2018].
[82] *IN RE FACEBOOK INTERNET TRACKING LITIGATION*. 2015.
[83] "Federal Trade Commission Staff Report: Self-Regulatory Principles For Online Behavioral Advertising: Tracking, Targeting, and Technology," *Federal Trade Commission*, 01-Feb-2009. [Online]. Available: https://www.ftc.gov/reports/federal-trade-commission-staff-report-self-regulatory-principles-online-behavioral. [Accessed: 10-Feb-2018].
[84] "Advertising Ethics | AAF." [Online]. Available: http://www.aaf.org/AAFMemberR/OUR_EFFORTS/Ethics/AAFMemberR/Efforts/Advertising_Ethcis.aspx?hkey=2e62934f-344e-473f-a7d2-30e6adad3229. [Accessed: 10-Feb-2018].
[85] "User Perceptions of Sharing, Advertising, and Tracking | USENIX." [Online]. Available: https://www.usenix.org/conference/soups2015/proceedings/presentation/chanchary. [Accessed: 08-Feb-2018].
[86] "Smart, useful, scary, creepy." [Online]. Available: https://dl.acm.org/citation.cfm?id=2335362. [Accessed: 09-Feb-2018].
[87] J. Turow, J. King, C. J. Hoofnagle, A. Bleakley, and M. Hennessy, "Americans Reject Tailored Advertising and Three Activities that Enable It," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 1478214, Sep. 2009.

[88] "What Tech Backlash? Google, Facebook Still Rank High in Polls | WIRED." [Online]. Available: https://www.wired.com/story/what-tech-backlash-google-facebook-still-rank-high-in-polls/. [Accessed: 09-Feb-2018].

[89] C. Newton, "How Americans really feel about Facebook, Apple, and more," *The Verge*, 27-Oct-2017. [Online]. Available: https://www.theverge.com/2017/10/27/16550640/verge-tech-survey-amazon-facebook-google-twitter-popularity. [Accessed: 09-Feb-2018].

[90] "2017 - Index Home | YouGov - BrandIndex." [Online]. Available: http://www.brandindex.com/ranking/2017-indexx. [Accessed: 09-Feb-2018].

[91] G. R. Milne, G. Pettinico, F. M. Hajjat, and E. Markos, "Information Sensitivity Typology: Mapping the Degree and Type of Risk Consumers Perceive in Personal Data Sharing," *J. Consum. Aff.*, vol. 51, no. 1, pp. 133–161, Mar. 2017.

[92] C. Mims, "Privacy Is Dead. Here's What Comes Next," *Wall Street Journal*, 06-May-2018.

APPENDIX A

EXTENSION MATERIALS

In this appendix, I include a variety of documents related to my live user data collection module. Specifically, I attach the following:

- Thesis overview document (provided to all study participants as a high-level outline of my methods and motivation)
- Participant consent form (template)
- Extension installation instructions
- IRB approval notice (#10183)

Leaking Ad Data: Thesis Overview

Online Privacy:
The online ad ecosystem is stunningly complex. Broadly, ad exchanges do two things:

- Follow you around the web to learn about your preferences and
- Allow advertisers to target advertisements based on your characteristics

There are tons of cookies that work on behalf of companies like Google, Facebook, Amazon, Appnexus, Rubicon, etc., that track the websites you visit and build user profiles based on this information. When you visit a website, they allow potential advertisers to 'bid' on serving an ad to you based on your unique background.

My Thesis:
My thesis investigates this ecosystem. My central undertaking is an attempt to understand whether we can reconstruct user profiles from targeted advertisement data. This work has two main implications: first, it will hopefully help researchers in the space understand more about how and why ads are targeted in particular ways, and the depth of information typically used in customization. Second, it will illustrate potential security risks that could arise if third parties are able to observe targeted ads shown to users.

This Extension:
This extension is the final step of the above work. Having used a simulated orchestration approach to train profile reconstruction models, I'll be using your anonymized data to see whether my program is able to rebuild the ad profiles of real, live users.

After two weeks of use, you will be eligible for:

1. Entry into a lottery for one of five $40 Amazon/Airbnb gift cards. To enter, you must submit your email address via the original popup form
2. A personalized report on ways to improve your privacy footprint, drawn from conclusions reached in my thesis. To access this anonymized report, you will need to save the userid string that will appear in the popup form after two weeks

Remember – personal identifiers will be removed from the data collected and all data will be kept confidential. Your name and other personal details (ie. section 1 of the popup form) will not be sent to my servers, and email addresses for the lottery are stored in a separate table (and not linkable to your ads or responses). All data is encrypted at rest and will be deleted at the end of the study. This study has received IRB approval (#10183).

Note: If you're looking for installation instructions, find them on the thesis popup.

ADULT CONSENT FORM
PRINCETON UNIVERSITY

TITLE OF RESEARCH: Leaking Data: Building User Profiles from Targeted Advertisements

PRINCIPAL INVESTIGATOR: *Edward Felten*

PRINCIPAL INVESTIGATOR'S DEPARTMENT: *Computer Science*

You are being invited to take part in a research study. Before you decide to participate in this study, it is important that you understand why the research is being done and what it will involve. Please take the time to read the following information carefully. Please ask the researcher if there is anything that is not clear or if you need more information.

Purpose of the research:
We wish to discover whether user profiles (ie. user interests, demographic information, and identifying information) can be reconstructed from targeted advertisements. Targeted user ads are incredibly prevalent in today's ad ecosystem, constructed based on ever more specific assessments of user interests. If third parties could reconstruct user profiles from targeted advertisements, serious questions would need to be asked about the stewardship of customer data by large organizations like Google or Facebook.

Study Procedures:
We will be collecting data on advertisements displayed to you through Google and Facebook, noting whether these ads are interest-targeted or more broadly location or context-targeted. From the information contained within these ads, we will attempt to derive a list of your interests. We will then benchmark our success against the interest lists identified for you by Google and Facebook, which we will ask you to self-report. Our program will also automatically note whether sensitive or personally-identifying data appeared in the advertisements shown to you – all such direct identifiers will be removed before we can access any of the advertisement data. You will be able to pause tracking at any point by temporarily disabling the extension or browsing in incognito mode, and stop tracking by uninstalling the extension. All direct identifiers will be removed and data will be kept confidential. Advertisement data will be tied only to a randomized study ID. All data will be destroyed at the end of the study.

Your total expected time commitment for this study is: *20 minutes to install the extension and complete a survey, 2 weeks with the extension installed (no minimum browsing requirement per day)*

Benefits and Risks:
Benefits: You will be permitted to opt-in to a drawing for one of five $40 gift cards. You will also be offered an opportunity to receive a report on your interest-categorizations. These reports will not contain any direct identifiers, and will be available to only those candidates who opt in. The reports can be an interesting way for you to learn more about the kinds of advertisements targeted to you.

Risks: The advertisement data collected from you may reveal information about user identity to the research team (ie. the ads themselves may directly reference identifying information). Browsing habits may be disclosed while the extension is in use (based on the kinds of advertisements displayed). You may also be slightly inconvenienced by having to disable adblocking extensions and having to see ads through the duration of the study

Alternatives
N/A

Confidentiality:
All records from this study will be kept confidential. Your responses will be kept private, and we will not include any information that will make it possible to identify you in any report we might publish. Research records will be stored securely in a locked cabinet, on password-protected computers, or on password-protected online databases. The research team will be the only party that will have access to your data.

Compensation:
Participants will be eligible to opt-in to a raffle for one of five $40 gift cards.

Who to contact with questions:

1. PRINCIPAL INVESTIGATOR:
   Edward Felten: felten@cs.princeton.edu

2. STUDENT RESEARCHER:
   Bharath Srivatsan: bharaths@princeton.edu

3. If you have questions regarding your rights as a research subject, or if problems arise which you do not feel you can discuss with the Investigator, please contact the Institutional Review Board at:

   Assistant Director, Research Integrity and Assurance
   Phone: (609) 258-8543
   Email: irb@princeton.edu

I understand the information that was presented and that:

A. My participation is voluntary, and I may withdraw my consent and discontinue participation in the project at any time. My refusal to participate will not result in any penalty.

B. I do not waive any legal rights or release Princeton University, its agents, or you from liability for negligence.

I hereby give my consent to be the subject of your research.

Click here to enter text.	Click here to enter text.
Subject's Signature	Date
Bharath Srivatsan	4/14/18
Person Obtaining Consent's Signature	Date

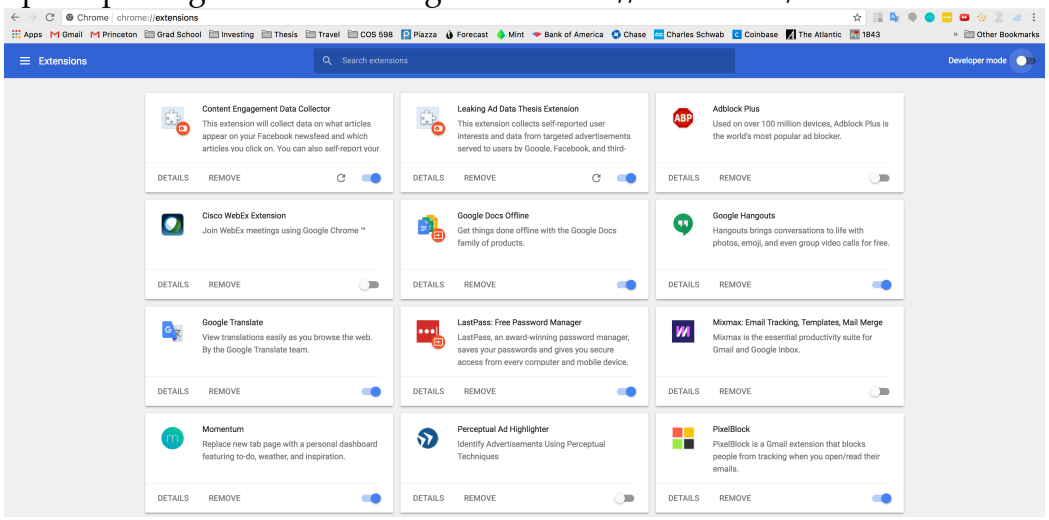
This study has been approved by the Institutional Review Board for Human Subjects

Ad Data Thesis Extension

Once again, THANK YOU for agreeing to install my extension and help me with my senior thesis! If you have any questions, please contact me at bharaths@princeton.edu.

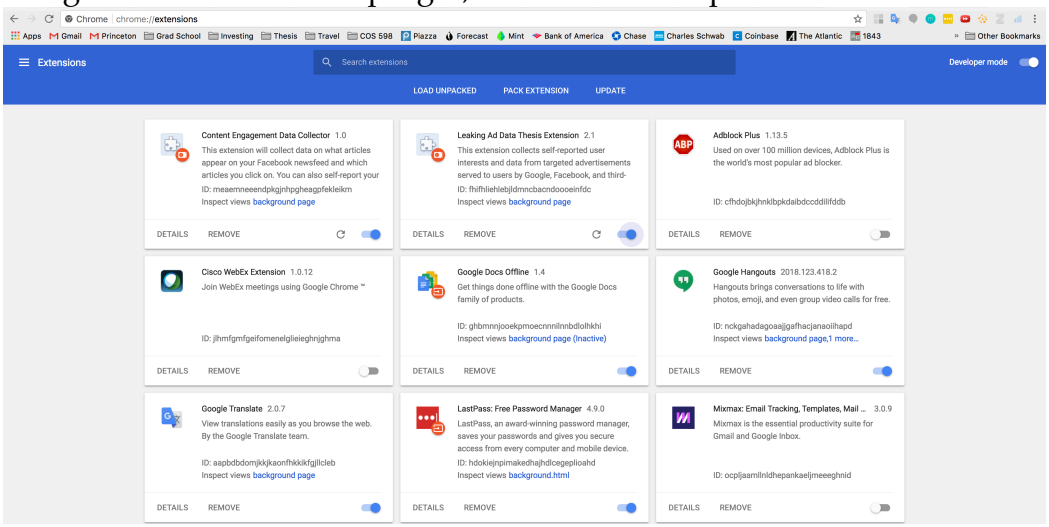
Participation steps:

1. Download, fill out, and send me (bharaths@princeton.edu) the consent form from https://drive.google.com/file/d/1BR-mbVoN41j5Qq2XcchbsU_1W0JAKUpc/view?usp=sharing
2. Download the extension folder from https://drive.google.com/file/d/1hFrJ_tItvhkCR-Ne6fwZexhYY8CTm3dT/view
3. Unzip the folder by double-clicking on it in your downloads directory
4. Open up Google Chrome and go to chrome://extensions/



Note: Your extension page may look different based on the extensions you have installed

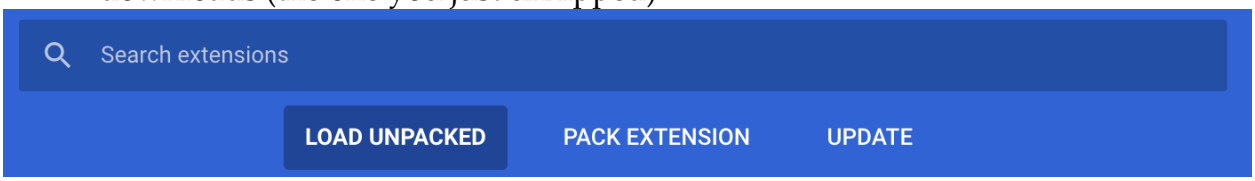
5. Using the button on the top right, turn on "Developer Mode"



This'll allow you to load extensions that aren't on the extension store

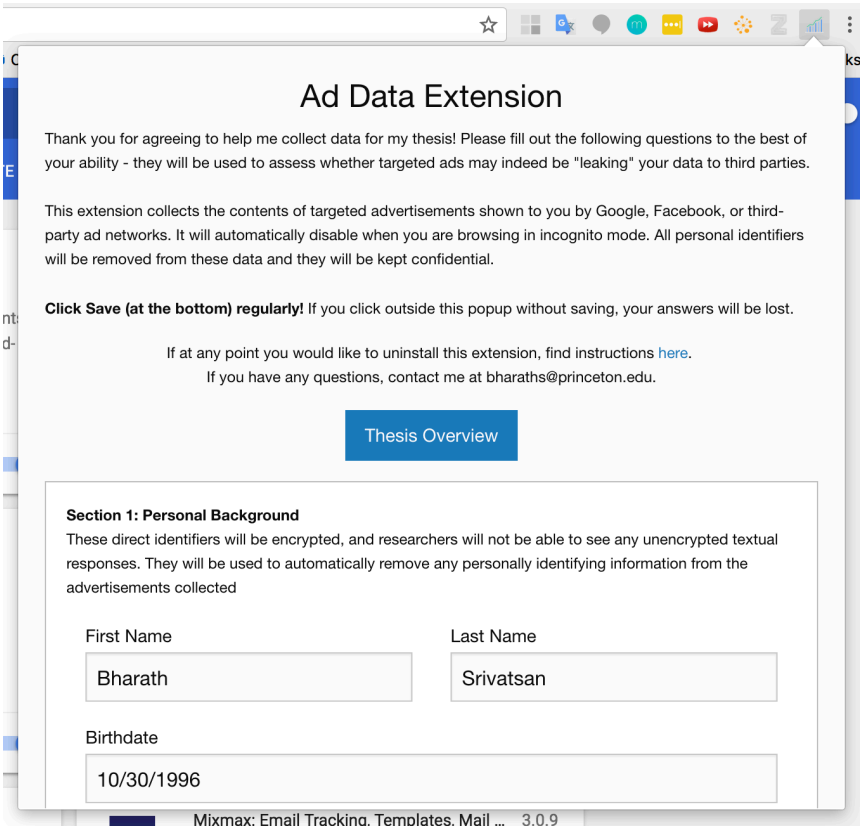
Important: If you have an adblocker installed, you must disable it for my extension to function. You can do so by switching the corresponding bar to 'off' (see the AdBlock Plus entry above)

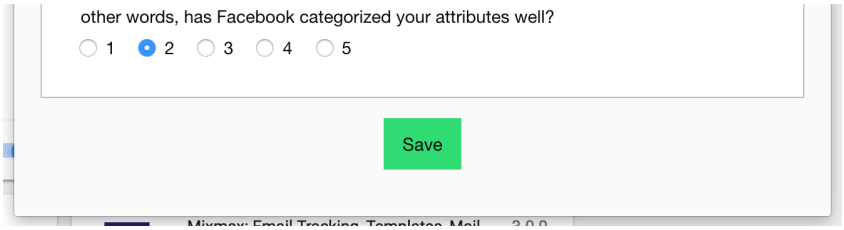
6. Click on "Load Unpacked", and select the "ad-data-extension" folder from your downloads (the one you just unzipped)



7. Click on the extension icon on the top right of your browser

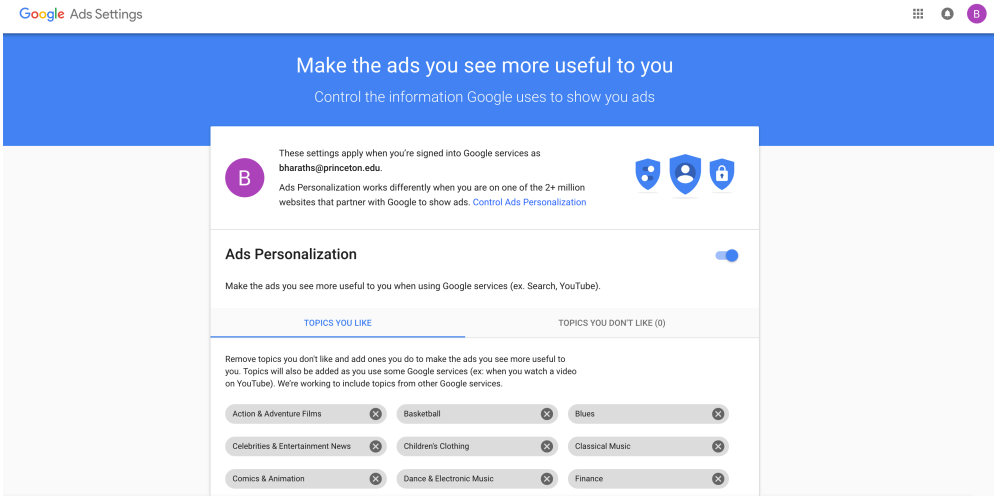
8. Fill out sections 1-3 to the best of your ability. **Note: Click save (on the bottom) regularly. If you click outside the popup without saving, your responses will be lost!**



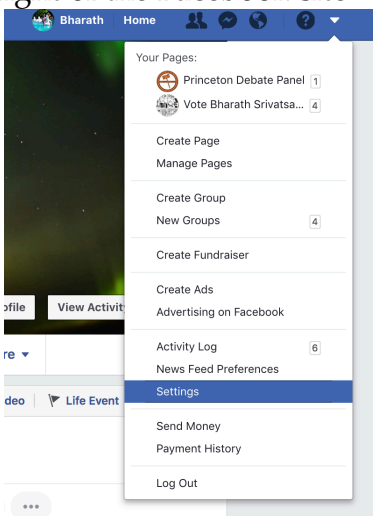


Click save regularly!

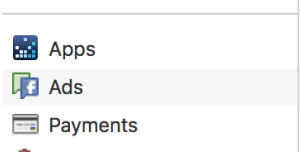
9. For section 4, find Google interests at https://www.google.com/settings/ads



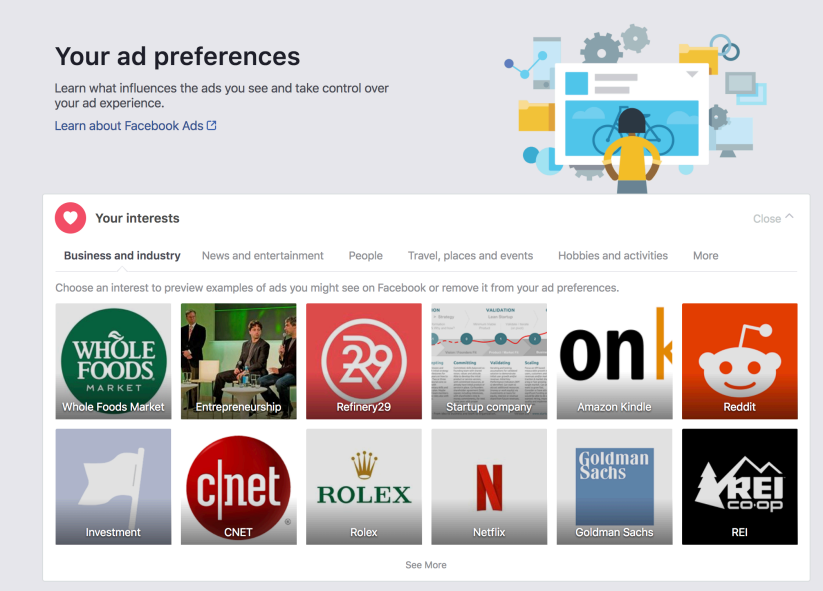
10. (click save!) Find Facebook ad preferences by clicking on settings from the drop down menu on the top right of the Facebook site



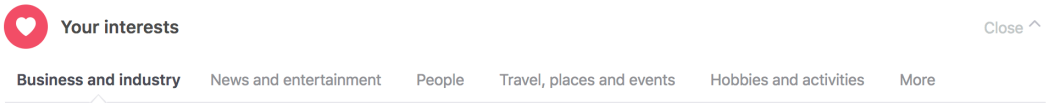
11. Then, click on "ads" on the left sidebar



12. On the ad preferences page, click on the "Interests" and "Your Information" dropdowns to see Facebook's assessment of your profile

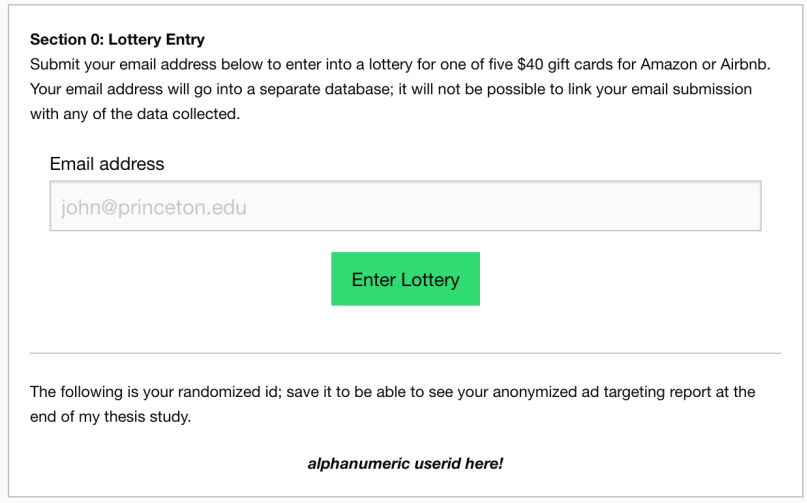


13. Copy interests from the categories featured on the main row (below: business, news, people, travel, and hobbies) and the 'categories' section of Your Information



14. Fill out section 5 to the best of your ability
15. Click save, and browse normally!

In two weeks, you'll be able to enter your email address to participate in the raffle. Save your userid to receive your personalized privacy report (instructions to follow).

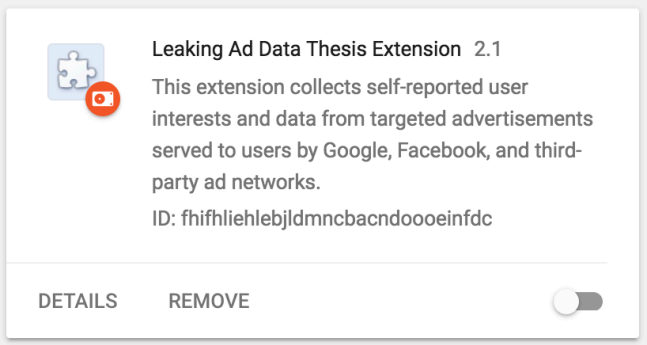


If for any reason you'd like to disable or uninstall my extension, follow the steps below.

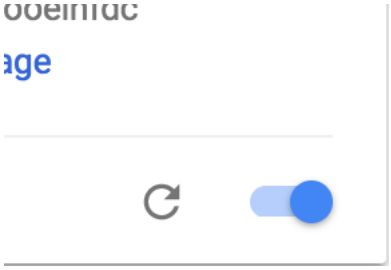
Note: The extension automatically switches off when you browse in incognito mode. You may disable and re-enable the extension as many times as you'd like. If you uninstall the extension before two weeks of use, you will not be able to enter the gift card raffle.

Deactivation steps:

1. Go to chrome://extensions/
2. Slide the blue bar to disable the extension temporarily. No further data will be collected while the extension is off.



3. When you'd like to re-activate, slide the blue bar back to "on"



Uninstallation steps:

1. Go to chrome://extensions/
2. Click on "REMOVE"

Research Integrity & Assurance
Princeton University
87 Prospect Ave., 3rd Floor
Princeton, NJ 08540

NOTICE OF APPROVAL

To: Felten, Edward William
From: Institutional Review Board
Re: IRB# 10183
Approved To: 14-Feb-2019

16-Feb-2018

Dear Edward Felten,

On 15-Feb-2018, the IRB approved the following study.

IRB#:	10183
Title:	Leaking Data: Building User Profiles from Targeted Advertisements
PI:	Felten, Edward William

Before the study's approval expires, you must secure approval to continue the study. This process is called continuing review. Note that that if the continuing review is not reviewed and approved by the approval end date, the study's approval will expire.

In conducting this study, you are required to follow the requirements in Princeton University IRB Policy #207: Obligations of the Principal Investigator for Human Subjects Research.

If you have any questions, please contact the IRB Office at (609) 258-0865 or irb@princeton.edu.

Thank you,

Edward P. Freeland, Ph.D.
IRB Chair

APPENDIX B

DATA COLLECTED

For each live user, I collected a series of advertisements and responses to a survey on ads, targeting and privacy. In this section I list these data points, outlining the ad objects I collected and my extension form questions.

All Ad Objects	`AdMetadata`	`AdContent`	`AdURLs`
Type: `goog/fb/…` Class : `pix/full/…` objid: `string` uid: `string` time: `int`	pii: `dict` url_pii: `dict`	text: `string` html: `string`	adName: `string` adUrls: `list` scriptUrls: `list` frameUrl: `string` parent: `string` adId: `string` adSrc: `string` linkText: `string` urls: `list`

Section 1: Personal Background
These direct identifiers will be encrypted, and researchers will not be able to see any unencrypted textual responses. They will be used to automatically remove any personally identifying information from the advertisements collected

First Name: Bharath
Last Name: Srivatsan
Birthdate: 10/30/1996
Place of Current Residence: Princeton | NJ | USA
Place of Home Residence: Singapore | State | Singapore

Section 2: Demographic Information
Please answer the following questions to the best of your ability. If you feel uncomfortable answering any of the below, leave the response field blank.

Age: 21
Gender: Male
Race (choose all that apply)
☐ Black and African American ☐ American-Indian or Alaskan Native ☑ Asian
☐ Hispanic or Latino ☐ Native Hawaiian or Other Pacific Islander ☐ White
Education Level: University - Undergraduate

Section 3: Privacy Practices
Please answer the following questions to the best of your ability. If you are unsure of your answer to any of the below, leave the response field blank.

Do you clear your cookies? Yes, once a year
Do you browse in incognito mode? Yes, occasionally
Do you often disable/block any of the following when browsing the web? (check all that apply):
☐ Cookies ☑ Javascript ☐ Location tracking ☐ Plugins/flash
Do you typically use an adblocker? ◉ Yes ○ No
Have you set up/elected to use the Do Not Track protocol? ○ Yes ◉ No
When you browse the web on Google Chrome, do you log in with your account? Yes, on all computers I regularly use

Section 4: User Profile
Instructions for how to complete this section are in the instructions document, also available here.

Google-identified user interests:
Blues
Business & Industrial
Business & Productivity Software
Business Finance
Celebrities & Entertainment News
Classical Music

On a scale of 1 (least) to 5 (most), how accurate would you say these interests are? In other words, are you interested in most or all of the topics Google found?
○ 1 ○ 2 ○ 3 ◉ 4 ○ 5

On a scale of 1 (least) to 5 (most), how comprehensive would you say these interests are? In other words, is Google missing out on interests that are significant to you?
○ 1 ○ 2 ◉ 3 ○ 4 ○ 5

Google-identified user profile
Gender: Male
Age: Unknown

Facebook-identified user interests:
TechCrunch
The Wall Street Journal
Gizmodo
Horror movies
Game of Thrones
Travel + Leisure

On a scale of 1 (least) to 5 (most), how accurate would you say these interests are? In other words, are you interested in most or all of the topics Facebook found?
○ 1 ○ 2 ◉ 3 ○ 4 ○ 5

On a scale of 1 (least) to 5 (most), how comprehensive would you say these interests are? In other words, is Facebook missing out on interests that are significant to you?
○ 1 ○ 2 ○ 3 ◉ 4 ○ 5

Facebook-identified user categories:
You do not have any behaviors in your ad preferences.

On a scale of 1 (least) to 5 (most), how accurate would you say these categories are? In other words, has Facebook categorized your attributes well?
◉ 1 ○ 2 ○ 3 ○ 4 ○ 5

Section 5: Outlooks on Online Privacy
Please answer the following questions to the best of your ability. Pick the closest descriptors, even if none fit you perfectly.

On a scale of 1 (least) to 5 (most), to what extent do you agree with the following statements?

When I browse the internet, I find that the ads I see are tailored to my interests
○ 1 ◉ 2 ○ 3 ○ 4 ○ 5
When I browse the internet, I want the ads I see to be tailored to my interests
○ 1 ○ 2 ◉ 3 ○ 4 ○ 5
I trust Google as a steward of my personal data
○ 1 ○ 2 ◉ 3 ○ 4 ○ 5
I trust Facebook as a steward of my personal data
○ 1 ◉ 2 ○ 3 ○ 4 ○ 5
I trust third-party advertisers as stewards of my personal data
◉ 1 ○ 2 ○ 3 ○ 4 ○ 5

Of the labels and categories assigned to you above, are any private? I don't mind Google/Facebook, but wouldn't want an advertiser seeing t
Having now seen your interest/user classifications, are you surprised? No, I'd expected such a profile
Having now seen your profile, do you plan to change how you browse? Yes, moderately

Extension survey screenshots (read as long page in two columns)

APPENDIX C

CODE

Project code: https://github.com/bsrivatsan/leaking-ad-data

The above repository includes the following code folders:

- `ad-data-extension`: All extension methods (see figure 5.6) and utility code. This folder is self-contained and can be run as-is (users installed the extension by loading this folder)
- `ad-data-processing:` A few processing scripts and utility methods used to analyze data. Includes a sample lambda trigger, a url-processing script to categorize ad topics, and a graphing script to visualize user connections. These files cannot be run as-is – they intentionally have been stripped of API keys and user data paths