Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers

Muhammad Ahmad Bashir†, Umar Farooq§, Maryam Shahid§, Muhammad Fareed Zaffar§, Christo Wilson†
†Northeastern University, §LUMS (Pakistan)
†{ahmad, cbw}@ccs.neu.edu, §{18100155, 18100048, fareed.zaffar}@lums.edu.pk

***Abstract*—Widely reported privacy issues concerning major online advertising platforms (e.g., Facebook) have heightened concerns among users about the data that is collected about them. However, while we have a comprehensive understanding *who* collects data on users, as well as *how* tracking is implemented, there is still a significant gap in our understanding: *what* information do advertisers actually infer about users, and is this information accurate?**

In this study, we leverage *Ad Preference Managers* (*APMs*) as a lens through which to address this gap. *APMs* are transparency tools offered by some advertising platforms that allow users to see the *interest profiles* that are constructed about them. We recruited 220 participants to install an IRB approved browser extension that collected their interest profiles from four *APMs* (Google, Facebook, Oracle BlueKai, and Neilsen eXelate), as well as behavioral and survey data. We use this data to analyze the size and correctness of interest profiles, compare their composition across the four platforms, and investigate the origins of the data underlying these profiles.

I. Introduction

If online advertising is the engine of the online economy, then data is the fuel for that engine. It is well known that users' behaviors are tracked as they browse the web [50], [27] and interact with smartphone apps [71], [70] to derive *interest profiles* that can be used to target behavioral advertising. Additional sources of data like search queries, social media profiles and "likes", and even offline data aggregated by brokers like Axciom [10], [48], are also leveraged by advertising platforms to expand their profiles of users, all in an effort to sell more expensive ads that are more specifically targeted.

Widely reported privacy issues concerning major advertising platforms [19], [18] have heightened concerns among users regarding their digital privacy. Several studies [55], [6], [84] have found that users are uncomfortable with pervasive tracking and the lack of transparency surrounding these practices. For example, a study by Pew found that 50% of online users are concerned about the amount of information that is available about them online [69].

Despite widespread concerns about online privacy and attention on the online advertising industry, there is one specific aspect of this ecosystem that remains opaque: what information is actually contained in advertising interest profiles, and are these profiles accurate? The extant literature presents a very complete picture of *who* is tracking users' online [7], [27], [9], [86], [71], [70], as well as *how* tracking is implemented (e.g., fingerprinting) [37], [78], [5], [57], [60], [72], [58], [1], [40], [28], but not necessarily the *what*. Controlled studies have shown that advertising platforms like Google do indeed draw inferences about users from tracking data [89], [20], [46], [47], but this does not address the broader question of what platforms actually know about users in practice.

Answering this question is critical, as it directly speaks to the motivations of the advertising industry and its opposition to privacy-preserving regulations. Specifically, the industry claims that tracking and data aggregation are necessary to present users with relevant, targeted advertising. However, if user interest profiles are incorrect, this would suggest that (1) online ads are being mistargeted and the money spent on them may be wasted, and (2) that greater privacy for users would have less of an impact on engagement with ads than is feared by industry. Further, surveys have found that users are deeply concerned when advertisers draw incorrect inferences about them [24], a situation that investigative journalists have anecdotally found to be true in practice [10], [48].

In this study, we leverage *Ad Preference Managers* (*APMs*) as a lens through which we answer these questions. *APMs* are transparency tools offered by some advertising platforms (e.g., Google and Facebook) that allow users to see the interest profiles that advertisers have constructed about them. Most of this data corresponds to user *interests* (e.g., "sports" or "computers"), although some corresponds to user demographics and attributes (e.g., "married" or "homeowner").

We recruited 220 participants to install an IRB approved browser extension that collected a variety of data (with user consent). First, it gathered the full interest profile from four *APMs*: Google, Facebook, Oracle BlueKai, and Neilsen eXelate. Second, it asked the participant to complete a survey on demographics, general online activity, and privacy-conscious behaviors (e.g., use of ad-blocking browser extensions). Third, the extension showed participants a subset of their interests from the *APMs*, and asked the participant if they were actually interested in these topics, as well as whether they recalled seeing online ads related to these topics. Finally, the extension collected the participant's recent browsing history and history of Google Search queries.

We use this mix of observational and qualitative data to investigate several aspects of the interest profiles collected by these four platforms, including: *How many interests do the APMs collect about users? Does each APM know the same interests about a given individual?* and *Are the interests (and*

Network and Distributed Systems Security (NDSS) Symposium 2019
24-27 February 2019, San Diego, CA, USA
ISBN 1-891562-55-X
https://dx.doi.org/10.14722/ndss.2019.23392
www.ndss-symposium.org

ads targeted to them) actually relevant to users? Further, by leveraging the last 100 days of historical data collected from our participants, we investigate what fraction of interests in the profiles could have been inferred from web tracking. We also use regression models to examine whether privacy-conscious behaviors and use of privacy tools have any impact on the size of interest profiles.

We make the following key contributions and observations throughout this study:

- We present the first large-scale study of user interest profiles that covers multiple platforms.
- We observe that interest profile sizes vary from zero to thousands of interests per user, with Facebook having the largest profiles by far. Because Facebook's profiles are so large, they typically have 59–76% overlap with the other platforms; in contrast, Google, BlueKai, and eXelate's profiles typically have $\leq$25% overlap for a given user, i.e., different platforms have very different "portraits" of users.
- Participants were strongly interested in only 27% of interests in their profiles. Further, participants reported that they did not find ads targeted to low-relevance interests to be useful. This finding suggests that many ads may be mistargeted, and highlights the need for a comprehensive study to evaluate the effectiveness of behavioral/targeted ads as compared to contextual/untargeted ads.
- We find that recent browsing history only explains $<$9% of the interests in participants' Facebook, BlueKai, and eXelate profiles (in the median case), while recent browsing and search history can only explain 45% of participants' Google profiles. Additionally, we find almost no significant correlations between privacy-conscious behaviors and interest profile size. These findings suggest that non-tracking data (data brokers) and other means of tracking like browser fingerprinting and cross-device tracking may be critical for building user interest profiles. It also suggests that alternate strategies and tools are needed to protect users' privacy.

Outline. Our study is structured as follows: we begin by providing background on *APMs* in § II. We describe our methodology and data collection in § III. In § IV, we present the results of our study. We survey related work in § V, discuss limitations of our study in § VI, and conclude in § VII.

II. Background

We begin by providing background information about the four companies that are the focus of this study. We manually visited the websites of dozens of top online advertising firms (based on data from prior work [8]) in Summer 2017 to determine which of them made transparency tools available to users. Only the following four provided functional tools at that time; the vast majority of companies did not offer an *APM*, and several were non-functional (e.g., Oath/Yahoo's).

Google. As of May 2018, Google is the largest online advertising company by revenue [30]. Google's empire includes ad networks and exchanges (e.g., DoubleClick) that place display ads within websites and mobile apps. Google is the largest purveyor of keyword-based ads via Google Search (i.e., AdWords) and video ads via YouTube.

Google's advertising products draw data from a variety of sources. Studies have consistently shown that resources from Google (e.g., Google Analytics and Tag Manager) are embedded on $>$80% of domains across the web [31], [13], [27], giving Google unprecedented visibility into users' browsing behavior. Google collects usage data from Chrome and Android, the world's most popular web browser and mobile operating system, respectively [87], [76]. Google's trove of user data includes real-time location (Android and Maps), text (Gmail, etc.), and voice transcripts (Assistant, etc.).

Google's *APM*[1] has been available since 2011 [64]. As of May 2018, this website allows a user to view lists of "topics" that the user does and does not like. The lists are primarily populated based on inferences from users' behavior, but users are able to add and remove topics from either list manually.

Table I shows frequent and random interests that we observed in Google's *APM* (the origins of our dataset are described in § III). We see that Google's inferred interests tend to be fairly broad, although some are more specific (e.g., "Cricket" instead of "Sports"). As we will see, other *APMs* present much more granular interests.

Google's policy states that they "do not show ads based on sensitive information or interests, such as those based on race, religion, sexual orientation, health, or sensitive financial categories" [35]. However, two audit studies have cast doubt on these claims, by demonstrating that Google did appear to infer sensitive attributes and allow advertisers to target them for ads [89], [20]. More troubling, these inferred interests did not appear in Google's *APM*.

Facebook. As of May 2018, Facebook is the second largest online advertiser by revenue [30]. Facebook primarily facilitates display and video advertising within its own products, including its social network, Instagram, etc. Facebook closed their public web ad exchange in 2016, but still operates an exchange for in-app mobile advertising [65].

Facebook's largest source of data are the personal profiles that its users curate about themselves. This includes static profile data (e.g., age and residence) and all of the Facebook Pages that users "like" and "follow". Facebook is also able to glean data from users' messages and observe web browsing behavior via its ubiquitous social widgets [75], [77], [16]. Until recently, Facebook partnered with data brokers like Acxiom and made the data accessible to advertisers [73].

Facebook's *APM*[2] includes several sections of information. One, named "Your Interests", lists all of the interests for the current user. Facebook provides brief explanations for the provenance of each interest, including liking a Page, or clicking on an ad related to the topic. Thus, interests on Facebook are drawn from a mix of explicit (liking) and implicit (clicking) data. Users may manually remove interests.

Another section, named "Advertisers You've Interacted With", lists companies and organizations that (1) have run

[1] https://adssettings.google.com/

[2] https://www.facebook.com/ads/preferences/

Platform	Top 10 Most Frequent Interests	10 Random Interests
Google	Computers and Electronics, Mobile and Wireless, Business and Industrial, Movies, Games, TV and Video, Finance, Celebrities and Entertainment News, Shopping, Travel	Fruit and Vegetables, Western Films, Photo Software, Travel Guides and Travelogues, Lawn Mowers, Reggaeton, Motorcycles, Combat Sports, Cricket, Fantasy Sports
Facebook Advertisers	Daraz Online Shopping, Airbnb, Spotify, The New Yorker, PayPal, Golf Digest, eBay, GQ, Target, VICE	Ralph Lauren, Marketing promotion, Beard Empire, OPPO, Gold's Gym, Waves Platform, HotPads, INSTAsmile UK, Capitol Cryo
Facebook Interests	Facebook Messenger, Facebook, Instant messaging, Social network, Instagram, Technology, Entre Rios Province, Food, Music, Education	Lensbaby, Laser lighting display, Toshiba, Mass communication, Scoop-Whoop, Steak sauce, Steam, Cricket For India, Mid-City New Orleans, International Federation of Accountants
BlueKai	Computer, English, Windows, The Academy Awards, Version 10.x (Windows 10), TV, Computers, Halloween Buyers, Pakistan, Summer Olympics Enthusiast	Outdoor Activities, Technology and Computing, Truck Rental, Web Analytics, Travel, Grocery Locations, Weekday Matinee, Length of Residence, Movies, Credit Cards
eXelate	Pets, Diet and Fitness, Tech - Enthusiasts, Shopping, Entertainment, Home and Garden, Hobbies, Finance, Auto Buyers, Finance and Insurance	Bed and Bath, Bluewave, Diet and Weight Loss, Pets, Lighting, Hobbies, Jewelry and Watches, Travel Enthusiasts, Auto Buyers - Sedan, Auto Buyers - Wagons, Finance and Insurance - Loans
ODP Categories	Music, Food, Television, News And Media, Movies, Clothing, Colleges And Universities, Sports, Shopping, Video Games	Strength Sports, Breaking News, Journals, Fantasy, Transportation And Logistics, Autos, Video, Reproductive Health, Beauty, Card Games

TABLE I: Examples of interests from each *APM*, advertisers from Facebook, and the ODP categories that we map all data into.

BlueKai Branded Data	Frequency
alliant	1624
acxiom	1392
datalogix	1266
acquireweb	1252
lotame	1145
affinity answers	1144
experian	1112
placeiq	739
adadvisor by neustar	658
tivo	620

TABLE II: BlueKai's top 10 data partners in our dataset.

ads on Facebook that (2) the current user has interacted with. However, Facebook's definition of "interaction" is extremely broad: it covers cases where the user has visited a company's website, used their app, clicked on one of their Facebook ads, or when the user was in a *Custom Audience* uploaded by the company (i.e., the company knew the user's name and email/phone number/zip code) [88]. Users may blacklist companies to prevent seeing their ads in the future.

As we show in Table I, interests on Facebook cover an enormous range of topics and granularities. Some are very general (e.g., "Education") while others are very specific (e.g., "Toshiba" and "Mid-city New Orleans"). We also show frequent and random advertisers on Facebook; since user must have interacted with these companies, we use them as proxies for user interests later in the study (e.g., visiting Ralph Lauren may indicate an interest in clothing and fashion).

Oracle BlueKai. BlueKai is a data broker that was purchased by Oracle in 2014 [62] and rebranded *Oracle Data Cloud*. BlueKai is not an advertiser; instead, they sell targeting data other third-parties. BlueKai collects data on users via cookie-based web tracking, and merges this information with data from partners such as banks (e.g., Mastercard), customer loyalty cards (e.g., DataLogix), and other data brokers.

BlueKai offers a website called *Registry*[3] that allows a user to see all of the interests and attributes that BlueKai has inferred about them. This information is looked up based on the user's BlueKai cookie (if it exists). Users cannot edit their inferred interests but they may opt-out of BlueKai.

As shown in Table I, BlueKai offers the most specific inferences of the companies we examine, including details about users' devices and home location. Much of this information is labeled as "branded data", meaning it originates from a data partner. Table II shows the top ten partners that appear in our BlueKai dataset, sorted by the amount of data they contribute. Degeling et al. provide further details about BlueKai's data ecosystem and products [22].

Nielsen eXelate. eXelate is a data broker that was purchased by Nielsen in 2015 [53]. Like BlueKai, eXelate's data originates from cookie-based web tracking, which is then linked to data from other sources. It is the smallest platform that we examine, in terms of the reach of its online trackers.

eXelate's *APM*[4] shows a user their interest profile by looking up their eXelate cookie (if it exists). Users may manually remove interests from their profile.

As shown in Table I, eXelate's inferred interests are the coarsest of the four platforms we examine. For the most part, eXelate's interests cover very broad topics (e.g., "Auto Buyers"), and only occasionally drill into more granular interests (e.g., "Sedans" and "Wagons").

III. Methodology

In this section, we introduce the datasets that we will use throughout this study. Ours is a complicated study that involves six separate datasets collected and processed using different methods; thus, we begin by presenting an overview of our datasets and their purposes.

A. Overview

The goal of our study is to examine the interests that major online advertising platforms have collected about online users. This includes both quantitative (how much do the platforms know?) and qualitative (are the interests correct?) analysis.

[3] http://bluekai.com/registry/

[4] http://exelate.com/privacy/opt-in-opt-out/

Furthermore, we aim to investigate possible origins of the interests, e.g., what about a users' behavior could have led to the drawing of a particular inference?

We require a number of different data sources to answer all of these questions:

- **Interest Profiles:** The key ingredient for our study are user interests profiles from online advertisers. We gathered interests from Google, Facebook, BlueKai, and eXelate's *APMs* by recruiting participants to install a browser extension. See § III-B.
- **Survey Data:** To understand how user demographics and behaviors (such as use of privacy-preserving tools) impact interest profiles, we asked our participants to complete a brief survey. We also asked participants to evaluate the quality of a subset of interests from their own profiles. See § III-B.
- **Browsing and Search History:** Our extension also collected participants' browsing history and Google Search history. We use this data to understand how users' behavior impacts the interests that are collected about them. See § III-B.
- **Third-party Trackers:** Although the browsing history data collected from our participants tells us the websites they have visited, it does not tell us which trackers observed them on each website. To bridge this gap, we crawled all 41,751 unique domains that appeared in our browsing history data. See § III-C.
- **Canonicalized Interests:** Directly comparing the interest profiles from different *APMs* is not possible because (1) they use different terminology, and (2) the interests vary widely in terms of specificity. To facilitate comparison of interests across *APMs*, we map them to a shared terminology space drawn from the Open Directory Project (ODP) [23]. See § III-D1.
- **Canonicalized Domains:** To compare participants' browsing and search history to their interest profiles (e.g., does visiting tennis.com lead to an inferred interest in "sports" or "tennis"?) we map 2^{nd}-level domain names to categories using a service provided by SimilarWeb [74]. See § III-D2.

In the following sections, we describe how all of this data was collected, cleaned, processed, and validated.

B. Collection of Interests & Survey Data

The most essential data for our study comes directly from web users, namely: the interests that online advertisers have collected about them; users' qualitative assessment of these interests; and user behavioral data from which these interests may be inferred. In this section, we explain how we collected and processed this data.

1) Browser Extension: We developed a browser extension that enabled us to collect interest profiles, survey data, and behavioral data from willing participants. Our extension is compatible with Google Chrome, and implemented the following process:

1) The extension opened a page in a new tab that explained our study, enumerated the data that would be collected, and asked for informed consent.

		Interests		
Platform	**Users**	**Unique**	**Total**	**Avg. per User**
Google	213	594	9013	42.3
FB Advertisers	190	6893	15392	81.0
FB Interests	208	25818	108930	523.7
BlueKai	220	3522	92926	422.4
eXelate	218	139	1941	8.9

TABLE III: Interests gathered from 220 participants, 82 from Pakistan and 138 from the US.

2) The extension checked whether the user was logged-in to Google and Facebook. If not, the user was asked to login, and the extension waited until they complied. Google and Facebook logins were necessary to collect interests from their respective *APMs*.
3) In the background, the extension collected a variety of data and sent it to our server. This included the user's browsing history (using the `chrome.history` API), Google Search history (from Google's *My Activity* website[5]), and all interests from the Google, Facebook, BlueKai, and eXelate *APMs*.
4) In the foreground, the extension asked the user to complete a survey.[6] The first portion of the survey contained static questions about demographics, general web usage, online activity (e.g., shopping for clothes, social media usage, etc.), privacy-conscious habits (e.g., clearing cookies), awareness of *APMs*, and usage of ad and tracker blocking browser extensions. Our demographic and privacy-related questions can be seen in Table IV and Table V.
5) The second portion of the survey included dynamically generated questions. Users were shown 20 randomly selected interests drawn from their Google, Facebook, and eXelate profiles[7] and asked for each one: whether the interest was relevant to them (5-point Likert scale), whether they had ever seen online ads related to this interest (Yes/No/Maybe), and if so, whether they found the ads to be relevant (Yes/No/Maybe).
6) The user was presented with a unique code to claim their remuneration, and told to uninstall the extension.

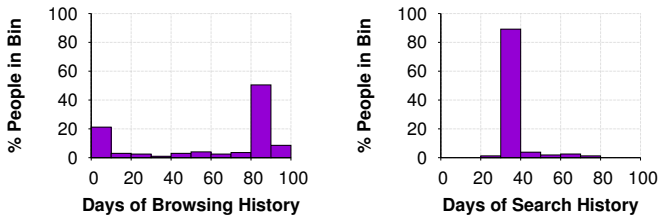
Overall, our survey took about 15 minutes to complete. The extension sent all collected data via HTTPS to a secure VM that was only accessible to the authors. We collected the most recent 100 days of browsing history and Google Search data from participants, to avoid downloading potentially enormous amounts of data. The Google Search data includes the keywords that participants' searched for and the URL they clicked on in the search results (if they clicked on something).

Our extension recorded all available interests from participants' Google, BlueKai, and eXelate interest profiles. Google and eXelate provide this data as text, whereas BlueKai presents images containing the interests (presumably to hinder crawl-

[5] https://myactivity.google.com/

[6] Our full survey instrument is available at https://cbw.sh/static/pdf/bashir-ndss18-survey.pdf.

[7] As we discuss in § III-B3, interests from BlueKai required offline processing that precluded them from the dynamic survey questions.



(a) Browsing history (b) Google Search history

Fig. 1: Amount of historical data collected from participants.

ing).[8] Our extension recorded the URLs to these images and we downloaded and parsed them offline (see § III-B3). From Facebook, our extension collected all interests shown in participants' "Your Interests" list and a subset of the advertisers in the "Advertisers You've Interacted With" list (see § II). Specifically, our extension recorded advertisers where the user had visited their website or used their app, because we view these as strong signals of interest on the part of the user.

With respect to the dynamic questions in our survey, the questions we asked were as follows. The first question was "Are you interested in X?", where X was an interest listed in one of the *APMs* for the current participant. This question did not concern a specific ad or purchasing intent; rather, the goal was to determine if the advertiser and the participant had matching expectations, i.e., do both agree that user U is generally interested in X? The second question was "Have you recently seen online advertisements related to X?", which narrows the context. The third question concerned intent: "Have the online ads you have seen related to X been relevant and useful to you?". Question three only appeared for a given interest if the user answered "Yes" to question two.

Ethics. Before we began data collection, we obtained IRB approval for our methodology from LUMS and Northeastern's review boards (protocols 2017-10-16 and #18-02-12, respectively). The first thing that the extension does after installation is disclose all information collection to the user and ask for consent. No data is collected until the user consents. All survey responses are anonymized.

2) Participant Recruitment: We recruited two groups of participants to our study: one from Pakistan, and one from the US. We did this to obtain a geographically diverse sample. Furthermore, we used different recruiting methods in each location to avoid potential biases in the resulting participant pool. In total, we recruited 220 participants: 82 from Pakistan and 138 from the US.

Participants from Pakistan were undergraduate students at LUMS. We advertised our study through university emails and student-focused Facebook groups. Participants were compensated with food coupons worth roughly $2.15. The vast majority of this data was collected from January 13–16, 2018.

US participants were recruited through Prolific [68], which is an online crowdsourcing platform similar to Amazon Mechanical Turk. Our tasks were confined to US Prolific users with high task completion rates (to discourage low-quality responses). Participants were recruited in two waves, with the first wave paid $5 for completing our survey, and the second paid $4.[9] This data was collected in March and May 2018.

To the best of our knowledge, no participants dropped out of the experiment due to a lack of Google or Facebook account, or because they did not use Chrome. These requirements were listed in our task description, so presumably people who did not meet them did not participate. Nine subjects refused to participate after reading our consent language, and three provided feedback: two had reservations about privacy, one "didn't have the time" to complete the task.

Table III presents an overview of the interest we were able to collect from our participants. Note that not all participants had data in all *APMs*; only BlueKai had interests about all 220 participants. Table IV presents the demographics of our participants. Figure 1 shows the amounts of historical data we were able to collect from participants. 50% of participants had 80–90 days of browsing history, while 20% had 0–10 days; this is possibly due to privacy-conscious users clearing their browsing history, which is a trait that participants reported in our survey. 90% of users had 30–40 days of search history.

3) Decoding BlueKai: Unlike the Google, Facebook, and eXelate *APMs*, which displayed user interests as text, BlueKai displayed interests as images (with one image per interest per user). Our extension collected all 43,420 URLs to these images from our participants, and we downloaded them offline.[10]

Once downloaded, we used the `Tesseract` Optical Character Recognition (OCR) [79] library to extract the text from each image. Unfortunately, we found that the text extracted by `Tesseract` was imperfect: using the `Enchant` library [45] we found that 79% of the OCRed words from our US participants were not valid English. These errors included misidentified letters and spacing anomalies (e.g., "Win dows" instead of "Windows"). To correct these OCR errors, we enlisted five undergraduate student volunteers to manually correct all of the text extracted from BlueKai images.

Validation. To verify that the manually curated text was correct, two of the authors independently and manually validated a random 10% sample of BlueKai interests by comparing the original images to the curated text. We found that the accuracy of the manually curated text was 95.3%.

C. Crawling Web Trackers

Recall from § III-B1 that our extension collected the browsing history of survey participants. The purpose of this data is to analyze whether users' browsing behavior influences their corresponding interest profiles. However, just because a user visits a specific website (say, tennis.com) does not necessarily mean that all four of the platforms will infer the user's interest (in this case, tennis). Only platforms that can observe users on a given site (i.e., by being directly embedded in the site, or by partnering with another third-party that is embedded [9]) can draw such an inference.

[8] On May 25, 2018, the day the GDPR went into effect, Google and BlueKai upgraded their *APMs* to present slightly richer data, and BlueKai began presenting information as text, not images.

[9] We reduced the reward after noting that recruitment was not an issue in wave one.

[10] We manually verified that the images and URLs shown to specific participants were accessible to us, and that the images were identical.

To bridge the gap between browsing histories and inferences drawn by specific platforms, we crawled the websites that participants had visited. Our dataset includes ∼1.2 million unique URLs from all participants. From these URLs, we extracted 41,751 fully-qualified domains. Using PhantomJS [66], we crawled all of these domains, as well as five randomly chosen links from each homepage that pointed to the same domain. The crawler recorded whether resources from the four *APMs* were included in each visited page. To increase the realism of the crawl, our crawler presented a valid `User-Agent`, scrolled pages, and waited 10–15 seconds on each page.

D. Canonicalization

The final step in our methodology concerns mapping the interests from each *APM* and participants' browsing and search histories into a single, canonical term-space, to facilitate comparison across *APMs*. We describe how we performed this mapping next.

1) Interests: One of the goals of our study is to compare users' interests profiles across platforms. However, there are several challenges that prevent direct comparisons. One issue is *synonyms*: *APMs* A and B may list a user's interests as "Real Estate" and "Property," respectively. A second issue is *specificity*: suppose that *APM*'s A and B, and C list a user's interests as "Sports," "Tennis," and "Wimbledon," respectively. All three have inferred the user's general interest in sports, but only C has drilled down to a specific tennis tournament.

To fairly compare the interest profiles of each platform, we must map the interests to a canonical term-space that has an appropriate level of specificity, i.e., not so general that all nuance disappears, but not so specific that interests like "Tennis" and "Wimbledon" fail to coalesce.

We investigated various ontologies, including DBpedia [21] and Wikifire [39], but ultimately settled on the Open Directory Project (ODP) [23] to supply our canonical term-space. ODP is a hierarchical categorization of websites that was manually curated by a volunteer team. Although websites are no longer being added to ODP as of 2017, the existing hierarchy of website categories provides an ideal term-space to canonicalize user interests (and online advertisers from Facebook). We chose to use the 465 categories[11] from the 2^{nd}-level of the ODP hierarchy as our common term-space. ODP only includes 15 1^{st}-level categories, which was too coarse for our use case, while we judged the 3,178 3^{rd}-level categories to be too broad.

To map our raw dataset to the 465 ODP categories, we had three people manually choose the single most appropriate ODP category for all 68K interests in our dataset. In cases of disagreement, the three labelers voted on the most appropriate category. Although we attempted to automate this process using techniques like Word2Vec [17], we found that the accuracy was low because (1) many of the "interests" in our data are actually specific companies (see Table I) which require background knowledge to appropriately map to a category, and (2) some of the interests are Pakistani words that are only accessible to native speakers.

[11]ODP actually includes 482 categories at the 2^{nd}-level, but we filtered out non-English terms.

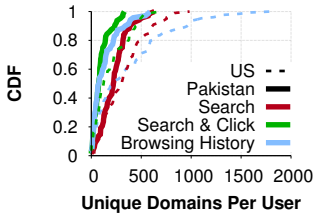


Fig. 2: Unique domains per participant in their browsing and search history. Solid (dashed) lines correspond to Pakistani (US) participants.

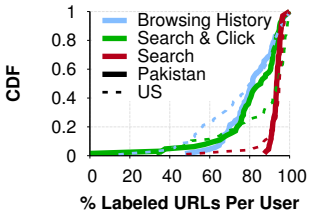


Fig. 3: Fraction of URLs per participant that were mapped to ODP categories. Solid (dashed) lines correspond to Pakistani (US) participants.

Validation. Similar to BlueKai, two of the authors independently and manually assigned ODP categories to a random 10% sample of interests. The authors' category labels agreed 93.8% of the time across all four *APMs*.

2) Domains: The last step in our methodology is mapping domains that our participants visited (recorded either in their browsing or Google Search history) to interest categories. For example, what interest would a tracker infer when a user visits tennis.com? Throughout this study, we use the term "domain" as shorthand for effective 2^{nd}-level domain names.[12]

Figure 2 shows the distribution of unique domains per participant in our study, broken down for Pakistani and US participants and by data source: browsing history from Chrome and search data from Google Search. "Search" refers to cases where a participant searched but did not click a resulting link, while "search & click" refers to cases where the participant clicked a search result. The latter is a strict subset of the former.[13] We observe that the US participants browsed and searched for more unique domains than the Pakistani participants. We also see that in both populations, "search" contributed more unique domains than browsing history.

We used SimilarWeb [74] to map domains to categories. SimilarWeb is a marketing tool that maps a given domain name to one of 221 categories. We manually crafted a 1:1 translation between the SimilarWeb and ODP categories, which was simple because they had obvious correspondence. Of the 51,541 unique domains in our dataset (covering browsing and search histories), SimilarWeb was able to categorize 77% of them. This classification rate is expected, given the long tail of websites that users visit in the real-world.

Figure 3 shows the fraction of URLs in participants' history that we were able to label using SimilarWeb (e.g., we would label both tennis.com/rackets and tennis.com/balls as "Sports"). We focus on individual URLs rather than domains to account for popularity, i.e., labeling a domain that is visited 100 times is more important than labeling a domain that was only visited once. We observe that the vast majority of URLs are successfully labeled, especially the URLs from Google Search. In the worst case, for browsing history of US

[12]Effective 2^{nd}-level domain names account for anomalous TLDs such as `.co.uk`.

[13]In the "search & click" case, we examine the domain that the participant clicked. In the "search" case, we examine the first domain that appeared in the search results, since we assume that it is representative of the information category that Google believed the participant was interested in.

	Pakistan	%	US	%
Gender				
Male	62	76	80	58.0
Female	20	24	54	39.1
Other	0	0	4	2.9
Age				
18-24	80	97.6	54	39.1
25-44	2	2.4	73	52.9
45-64	0	0.0	11	8.0
65 plus	0	0.0	0	0.0
Educational Attainment				
None	0	0.0	0	0.0
High School	32	39.0	41	29.7
College	37	45.1	62	44.9
Some graduate school	9	11.0	16	11.6
Masters	4	4.9	16	11.6
Doctoral	0	0.0	3	2.2
Residence Location				
Urban	67	81.7	77	55.8
Suburban	13	15.9	54	39.1
Rural	2	2.4	7	5.1
Marital Status				
Never married	75	91.5	96	69.6
Married	3	3.7	33	23.9
Divorced	0	0.0	6	4.3
Separated	0	0.0	1	0.7
Widowed	0	0.0	0	0.0
I prefer not to say	4	4.9	2	1.4
Number of Children				
0	68	82.9	98	71.0
1	2	2.4	24	17.4
2	8	9.8	11	8.0
3	3	3.7	2	1.4
4	1	1.2	3	2.2
5 or more	0	0.0	0	0.0
Employement				
Yes, Full-time	6	7.3	68	49.3
Yes, Part-time	8	9.8	28	20.3
No	68	82.9	42	30.4

TABLE IV: Participant demographics.

users, we were still able to label more than 60% of URLs for 77% of participants. As expected, browsing history was the most challenging dataset to label, since users may visit unpopular websites (whereas Google Search tends to show links to popular websites).

Validation. Once again, two of the authors independently and manually validated the categorization of a 10% random sample of domains. The authors rated the accuracy of the SimilarWeb categories at 98%.

IV. Analysis

Having introduced our various datasets, their provenance, and validation, we now move on to analysis. *First*, we briefly discuss the salient characteristics of our participants. *Second*, we compare and contrast the size and composition of interest profiles across the four *APMs*. *Third*, we examine whether the inferences drawn about users by the platforms are accurate, based on the responses to our dynamic survey questions. *Finally*, we investigate how users' demographics and online activity may relate to their interest profiles.

A. Participants

We begin by examining the demographics of participants in our sample as shown in Table IV. We see that the characteristics of our populations match our recruitment strategies:

	Pakistan	%	US	%
How Often Do You Clear Your Browsing History?				
Never	41	50.0	45	32.6
Monthly	28	34.1	63	45.7
Weekly	10	12.2	21	15.2
Daily	1	1.2	6	4.3
Multiple times a day	1	1.2	3	2.2
How Often Do You Clear Your Cookies?				
Never	44	53.7	38	27.5
Monthly	30	36.6	67	48.6
Weekly	8	9.8	28	20.3
Daily	0	0.0	4	2.9
Multiple times a day	0	0.0	1	0.7
How Often Do You Browse in Private Mode?				
Never	20	24.4	37	26.8
Monthly	21	25.6	33	23.9
Weekly	25	30.5	46	33.3
Daily	8	9.8	16	11.6
Multiple times a day	8	9.8	6	4.3
Are You Aware of eXelate?				
Yes	5	6.1	3	2.2
No	77	93.9	135	97.8
Are You Aware of BlueKai?				
Yes	4	4.9	4	2.9
No	78	95.1	134	97.1
Have You Ever Opted-out of Online Advertising?				
Yes	24	29.3	72	52.2
No	35	42.7	42	30.4
I don't know	23	28.0	24	17.4
Do You Have Do Not Track Enabled in Your Browser?				
Yes	12	14.6	33	23.9
No	33	40.2	56	40.6
I don't know	37	45.1	49	35.5
Do You Have the AdBlock Extension Installed?				
Yes	41	50.0	53	38.4
No	30	36.6	79	57.2
I don't know	11	13.4	6	4.3
Do You Have the AdBlock Plus Extension Installed?				
Yes	18	22.0	41	29.7
No	45	54.9	91	65.9
I don't know	19	23.2	6	4.3
Do You Have the uBlock Origin Extension Installed?				
Yes	4	4.9	36	26.1
No	53	64.6	98	71.0
I don't know	25	30.5	4	2.9
Do You Have the Ghostery Extension Installed?				
Yes	2	2.4	15	10.9
No	54	65.9	119	86.2
I don't know	26	31.7	4	2.9
Do You Have the Disconnect Extension Installed?				
Yes	4	4.9	5	3.6
No	54	65.9	129	93.5
I don't know	24	29.3	4	2.9
Do You Have the Privacy Badger Extension Installed?				
Yes	3	3.7	11	8.0
No	54	65.9	123	89.1
I don't know	25	30.5	4	2.9

TABLE V: Online privacy and advertising preferences of participants.

the Pakistani participants are younger, more urban, unmarried, have few if any children, and are mostly unemployed (i.e., they are full-time students), compared to our US participants. Overall, men are over-represented in our population (65% of all participants), while older individuals (45+ years old) are under-represented (5% of all participants).

Table V shows the responses to the privacy-focused questions in our survey. The Pakistani participants are somewhat less privacy conscious than the US participants: more than 50% reported never clearing their browsing history or

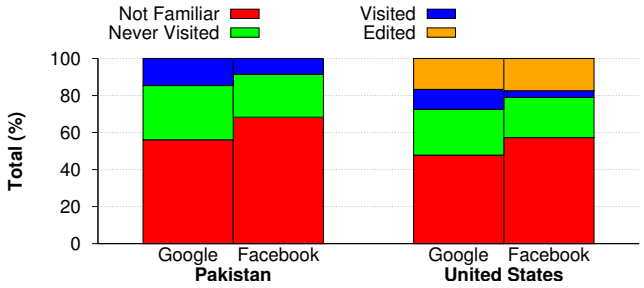


Fig. 4: Participants' awareness of Google and Facebook *APMs*.

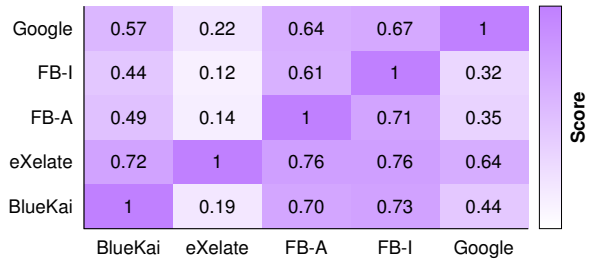
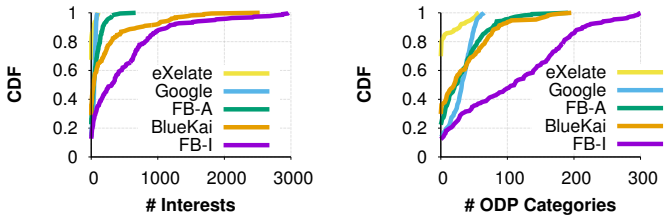


Fig. 6: Overlap of ODP-categorized interests across platforms.



(a) Raw interests. (b) ODP categorized interests.

Fig. 5: CDF of interests per user for all platforms.

cookies, versus 28–33% of US participants; only 29% of Pakistani participants reported opting-out of some form of online advertising, versus 52% of US participants; and the average number of privacy-preserving browser extensions per Pakistani participant was 0.88, versus 1.17 for US participants. Pakistani participants also expressed less awareness of privacy-preserving browser extensions, with roughly 30% responding "I don't know" if one was installed in their browser, versus 3% for US participants.

There were areas of agreement between our participants pools with respect to privacy. Both Pakistani and US participants reported using Private Mode in their browser at similar rates (24–26% reported using it monthly, while 31–33% reported using it weekly). AdBlock [2] (not to be confused with AdBlock Plus [32]) was the most popular privacy-preserving browser extension in both cohorts, although it was more popular with Pakistani participants. Finally, roughly 40% of both cohorts reported not having Do Not Track (DNT) enabled in their browser. The DNT question also elicited the greatest amount of confusion among all our questions, with 36–45% of participants answering "I don't know."

Another area of agreement between our participants was their lack of awareness of tracking companies and *APMs*. Over 90% of participants reported that they had never heard of BlueKai or eXelate, which is not surprising since they are not consumer-facing companies. As shown in Figure 4, 48–68% of participants were unaware of Google and Facebook's *APMs*, and these numbers rise to 72–91% if we include participants who reported awareness of the *APMs* but never visited them.

One surprising finding, shown in Figure 4, is that 17% of US participants reported that they had edited their interests in Google and Facebook's *APMs*, versus zero Pakistani users. We will return to this observation in § IV-C when we investigate whether the interests in the *APMs* are relevant to users.

B. Interest Profiles

Next, we move on to analyzing the interests in participants' *APM* profiles. Table III shows the summary statistics for interests in our dataset. Recall that Facebook's *APM* displays interests (drawn from explicit "likes" and inferred from user behavior) and advertisers that users have interacted with (see § II); we separate these two datasets in our analysis.

As shown in Table III, only BlueKai had an interest profile for all participants. Overall, Facebook's interest profiles contained the most information (524 interests per user on average) followed by BlueKai (422 interests per user), while eXelate had the least (nine interests per user). Across the board, we find a one order of magnitude reduction in interests when we remove duplicates, indicating large amounts of overlap between users' interests.

Figure 5a shows the distribution of interest profile sizes for our participants. Facebook interest profiles are hands-down the largest, with 38% of participants having 500+ interests. 13% of our participants also had BlueKai profiles with 500+ interests, whereas Google and eXelate profiles never contained more than 92 and 107 interests, respectively.

However, it is possible that these comparisons of profile size are misleading: perhaps the Facebook and BlueKai profiles contain hundreds of very specific, synonymous interests (e.g., tennis, table tennis, and paddle ball), whereas the Google and eXelate profiles only contain high-level, non-overlapping interests (e.g., sports). To make the comparisons between platforms fair, we map all interests into a shared term-space of 465 ODP categories (see § III-D1) and present the distribution of profile sizes after canonicalization in Figure 5b. As expected, this process reduces the size of interest profiles overall, but the relative order of the five distributions does not change **at all** from Figure 5a to Figure 5b. This proves that the Facebook interest profiles in our dataset do indeed contain the greatest variety of interests per user. BlueKai, on the other hand, loses diversity: in Figure 5a the BlueKai profile sizes handily outpace Facebook advertisers, but in Figure 5b the two distributions are close together.

Overlap. Next, we examine the overlap between interests on different platforms using the canonicalized interests that we mapped to ODP categories.

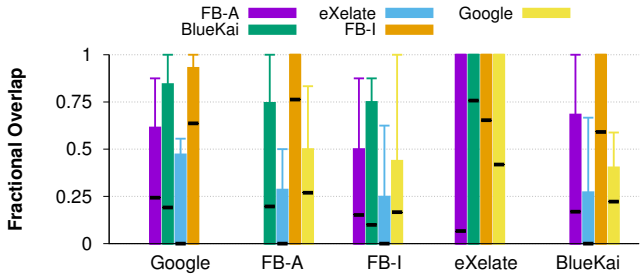


Fig. 7: Per participant overlap of ODP-categorized interests across platforms. Each box-and-whisker shows the minimum, 5^{th} percentile, median, 95^{th} percentile, and maximum overlap between user profiles for the given pair of *APMs*.

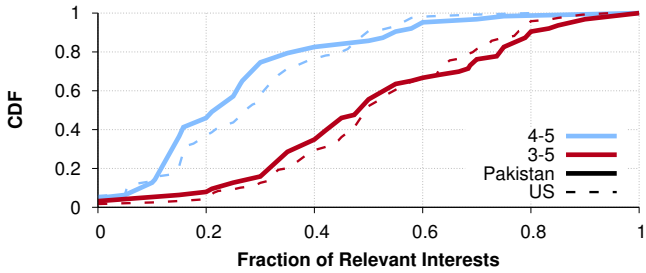


Fig. 8: Fraction of interests rated as relevant by participants. Solid (dashed) lines denote Pakistani (US) participants.

Figure 6 presents the overall overlap between the set of interests we observe on each platform. Each square in the heatmap is calculated as $\frac{|X \cap Y|}{|Y|}$, where X and Y are the sets of interests from the *APMs* on the x- and y-axis of the figure. The results are intuitive: the smallest *APMs* in terms of number of unique interests (eXelate and Google) are well-covered by the larger *APMs* (Facebook and BlueKai), with overlaps in the 0.57–0.76 range. Conversely, the smaller *APMs* fail to cover their larger rivals, with overlaps in the 0.12–0.44 range.

While Figure 6 gives an idea of how well each *APM* covers the entire space of interests, it does not reveal whether the interest profiles for individual users have overlap across *APMs*. To answer this question, we plot Figure 7, which shows the distribution of per user interest overlap between pairs of *APMs*. Each box-and-whisker shows the minimum, 5^{th} percentile, median, 95^{th} percentile, and maximum overlap between user profiles for the given pair of *APMs*. For example, the left-most, purple box shows that the median Google user's interest profile has 24% overlap with their Facebook advertisers profile.

In the majority of cases, the median overlap between pairs of *APMs* is $\leq$25%, meaning that different platforms are drawing substantially different inferences about users. However, there are two exceptions. *First*, the median Facebook interest profile covers 59–76% of the interest profiles of the other four *APMs* (including Facebook advertisers), demonstrating that Facebook's coverage of interests tends to subsume the other platforms that we study. *Second*, eXelate, being the smallest platform in our study, has low coverage of the others (median: zero), and conversely is well covered on average by the others.

C. Perceptions of Interests & Ads

Thus far, we have found stark differences between the *APMs* we have examined, both in terms of the size of the interest profiles, as well as the lack of overlap between them. This heterogeneity raises our next set of questions: are some *APMs* able to collect more *relevant* interests than others? In other words, when an *APM* draws an inference about a user's interest, is this inference correct? This question strikes at the heart of the targeted advertising industry: if the features used for ad targeting are not actually relevant to users, then the money spent placing those ads may be wasted.

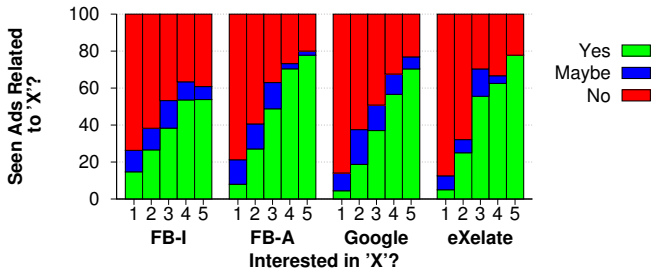
To address this question, we showed each participant in our study 20 randomly selected interests from their interest profiles and asked them to rate how relevant they were on a 5-point Likert scale ranging from "very" to "not at all" (see § III-B). Interests were selected proportionally to the size of each participant's *APMs*, with a maximum of 12 interests from any single *APM*.[14] Because BlueKai presented interests as images, we could not dynamically show them to participants, thus we omit analysis of BlueKai from this section. We also asked participants whether they recalled seeing ads that were related to each interest (Yes/No/Maybe), and if so, whether each of those ads was "relevant or useful" (Yes/No/Maybe).[15]

Figure 8 shows a CDF of the fraction of interests that participants judged to be relevant. The lines capture different thresholds for relevance: one considers interests in the range $[4, 5]$ as relevant, while the other includes interests in the range $[3, 5]$. In the more permissive case, 52–56% of participants said less than half of their interests were relevant; in the stricter case, 86-91% of participants said the same thing. These results are consistent across our two participants pools. This is somewhat surprising, since 17% of our US participants reported editing their Google and Facebook interest profiles (see Figure 4), presumably by deleting irrelevant and incorrect interests. In particular, the distribution (not shown) of these 17% US participants is consistent with that of Pakistani participants who did not edit their interest profiles. We further note that our results for interest relevance are consistent across the *APMs*: no platform contained more relevant interests than the others. This suggests that (1) our US users may have over-reported the rate at which they edited their interest profiles, (2) perhaps participants only edited their profiles slightly, or (3) the *APMs* added additional (mostly irrelevant) interests to participants' profiles after the manual editing took place.

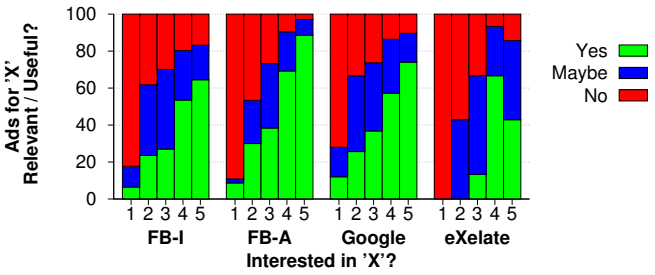
Next, we examine the relationship between interest relevance and the likelihood of seeing related online ads. Figure 9a shows the fraction of interests where participants did or did not recall seeing a related ad, broken down by the relevance of the interest and by *APM*. Our first observation is a general trend across all four *APMs* that participants remembered seeing more ads for more relevant interests. One potential explanation is that people are prone to remembering relevant advertising. Another possibility is that advertisers are able to determine which of a given person's interests are "real" (possibly by measuring engagement), and thus target more ads to more relevant interests.

[14]This design ensured that each participant was presented with 20 interests, even if one of their profiles was sparse. The maximum ensured diverse selection across the four *APMs*, and primarily impacted interests from Facebook.

[15]Please refer to § III-B1 and our survey instrument for more details.



(a) Interest Relevance vs. Seeing Ads



(b) Interest Relevance vs. Seeing Relevant Ads

Fig. 9: Participants' rating of interests (5-point Likert scale with 1 being "not at all") versus (a) whether they recall seeing ads related to that interest and (b) whether ads related to that interest were relevant or useful. Note that we do not have survey data for BlueKai because it was not possible to scrape interests from their website in real-time (see § III-B3).

Our second observation from Figure 9a is that participants recalled seeing similar distributions of ads across the interest relevancy scale for eXelate, Google, and Facebook advertisers. In contrast, users recalled seeing more ads linked to irrelevant Facebook interests. One potential explanation for this stems from our prior observation that Facebook interest profiles are by far the largest in our study (see Figure 5). Facebook appears to allow their interest profiles to grow indefinitely; conversely, Google and eXelate may be more aggressive about pruning profiles to eliminate stale interests. Thus, the breadth of Facebook's interest profiles may make it more difficult for advertisers to target users' relevant interests.

Now, an advertiser might contend that users are not the best judges of interest relevance. For example, a person might not realize they are interested in a new car until they see an ad about one. In other words, it is possible that users may judge a specific ad to be useful, even if they judge the underlying interest as irrelevant.

To investigate this further, we plot Figure 9b, which shows the fraction of interests where participants judged the related ads as useful or relevant, *conditioned on the participant seeing an ad.* In other words, Figure 9b focuses on the subset of "Yes" cases from Figure 9a. We observe a clear trend where ads related to relevant interests were themselves judged to be relevant. Overall, our participants reported that 49% of the ads they saw were linked to highly relevant (scores $[4, 5]$) interests. If people have a retrospective bias that prevents them from recalling irrelevant ads (or ads related to irrelevant interests), this means our results **underestimate** the volume of mistargeted ads.

D. Origins of Interests

The next question we investigate is: to what extent do users' demographics and behavior impact their interest profiles? Prior controlled studies have proven that the websites visited by users impact the interests that appear in Google's *APM* [89], [20]. Although it is likely that browsing history also impacts the interest profiles constructed by Facebook, BlueKai, and eXelate, to the best of our knowledge this has not been conclusively demonstrated. Furthermore, it is unclear what fraction of interests in users' profiles are inferred from browsing history, as opposed to other sources of data like search history (in the case of Google), smartphone telemetry, data partners (see Table II), etc.

Fig. 10: Overlap of participant history with each of the *APMs*. Each box-and-whisker shows the minimum, 5^{th} percentile, median, 95^{th} percentile, and maximum overlap.

To investigate this question, we analyze the overlap between our participants' recent browsing and search history and the interests in their *APM* profiles. Essentially, we pose the question: *what fraction of interests in a participant's profile could possibly be explained by their recent online activity?* For this analysis, all interests and domains from historical data are canonicalized to ODP categories to make them comparable.

Figure 10 shows distribution of overlaps between participants' historical data and their interest profiles (normalized to the size of their interest profile) on the five *APMs*. We compare participants' browsing history to all five *APMs*, since all five companies are known to track web users. For a given *APM* A, we only consider domains in participants' browsing history that include trackers from A, based on our crawls of these domains (see § III-C). We only compare participants' search history (including searches that did and did not result in clicks on links) to Google's *APM*, since only Google has access to search history. The right-most box-and-whisker shows the overlap between the union of participants' recent browsing and search history with their Google interest profiles, i.e., it combines several data sources that are all available to Google.

With respect to Facebook, BlueKai, and eXelate, we find that participants' recent browsing history does a poor job of covering the interests in their profiles. On eXelate and

Feature	Dependent Variable: Number of Interests per User				
	Google	**FB-A**	**FB-I**	**eXelate**	**BlueKai**
Intercept	**2.9142*****	**3.5242*****	**5.2361*****	**3.8091*****	**3.8091*****
Demographics	✓	✓	✓	✓	✓
Time Spent Online Per Day	-0.0994	-0.5605	-0.2493	0.0097	0.0097
Web Searches Per Day	0.4156*	0.0045	-0.4014	1.5893*	1.5893*
Size of Browsing History	3.964e-06	3.519e-06	-1.087e-05	**5.997e-05*****	**5.997e-05*****
Size of Google Search History	**7.16e-05*****	5.661e-05*	1.793e-05	-7.402e-06	-7.402e-06
Posting on Social Media	-0.1624	-0.0842	0.7758*	0.4066	0.4066
Reading Social Media	0.1685	0.9426*	1.1643**	0.0131	0.0131
Interactions With Online Ads	✓	✓	✓	✓	✓
Interactions With Search Ads	✓	✓	✓	✓	✓
# of Ad Block Extensions	0.0480	0.0300	-0.0363	-0.1453	-0.1453
Use of a Privacy VPN	0.0470	0.3770	0.2259	0.3828	0.3828
How Often Do You Clear Cookies	-0.4799	0.4087	-0.1976	1.1986	1.1986
How Often Do You Clear Browsing History	0.5593*	0.0217	-0.2365	0.9350	0.9350
How Often Do You Browse in Private Mode	0.0760	-0.6484*	-0.2940	-0.8539	-0.8539
Use of DNT	**-0.6375*****	-0.5022*	-0.5204*	-0.5432	-0.5432
Awareness of APMs	✓	✓	✓	✓	✓
Observations:	213	190	208	220	218
Log-Likelihood:	-1012.9	-1124.2	-1501.5	-456.86	-1402.9
Pearson χ^2:	99.3	371	268	387	740

TABLE VI: Estimated coefficients of negative binomial regressions on the size of *APM* profiles. Italicized rows indicate groups of related, independent features that we control for in the models, but do not investigate for significance. *Note:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

BlueKai, the overlap of our median participant's interest profile with domains they browsed is almost zero. For Facebook advertisers, the overlap is only 9%.

With respect to Google, we find that recent browsing history does a substantially better job of covering interests in our participants' profiles, with 30% coverage in the median case. There are three possible explanations for this. *First*, Google may be more reliant on tracking data than the other platforms that we study (Google does not have a major social network like Facebook, nor are they known to partner with data brokers like Facebook and BlueKai). *Second*, Google's trackers have far greater coverage of the web than any other company [27], possibly suggesting that they simply have access to more and better data from which to draw inferences. *Third*, Google may be more aggressive than the other platforms about pruning old interests in the profiles and replacing them based on recent activity; conversely, the other platforms may be more reliant on older historical behavior that is not captured in our dataset.

We observe that the domains that our participants click on in Google Search results cover roughly the same fraction of interests as browsing history, although expanding the set to include searches that did not result in a click increases coverage in the median case to 41% and reduces variance. This result is particularly interesting because our search history data is more complete than our browsing history data: Google Search history aggregates data from all devices where a participant is logged-in to their Google account (e.g., desktop and mobile devices), while our browsing history data is only drawn from the participant's desktop browser. Thus, it is surprising to see that combining all three data sources only covers 45% of interests for our median participant. This suggests that Google, like the three platforms, also infers user interests from sources that are not covered by our dataset.

Regression. Next, we expand our investigation of factors that may impact participants' interest profiles, including: demographics (see Table IV); general use of the internet, web search, and social networks; and privacy-preserving behaviors (see Table V). We also include measured data such as the size of participants' recent browsing and search history.

Our goal is to understand whether any of these features impact the size of participants' interest profiles, i.e., are there specific classes of people, or specific behaviors, that correlate with smaller interest profiles? To investigate these relationships we use regression tests. We adopt a negative binomial model and regress on the features of individual participants, i.e., the independent features are a matrix of the survey and historical data collected from our browser extension, and the dependent feature is a vector of interest profile sizes. We fit five models, one per *APM*. We chose to use a negative binomial model since it is appropriate for a count dependent variable, and it handles over-dispersion in the data.

Table VI shows the coefficients and p-values for our five fit models. Negative coefficients indicate negative correlation between a given feature and the size of interest profiles. Note that the independent features are of different magnitudes, meaning that only sign and significance can be compared across features. For survey questions that included an ambiguous reply (e.g., "Maybe" or "I don't know"), we conservatively map these responses to "No." Italicized rows in Table VI indicate groups of related, independent features that we control for in the models, but do not investigate for significance.[16] The largest Variance Inflation Factor (VIF) we observed for any feature in any of our models is 2.7, which indicates that our independent features do not suffer from multicollinearity.[17]

With regards to general use of the internet and online services, we observe three strong ($p < 0.001$) effects. *First* and unsurprisingly, engagement with social media is positively correlated with Facebook interest and advertiser profile length. *Second*, the size of participants' browsing and search histories

[16] In total, these rows contain 20 features.

[17] Features with VIF scores $\geq$5 are typically eliminated from models to remove multicollinearity.

are strongly, positively correlated with interest profile sizes on eXelate and BlueKai in the former case, and Google in the latter. Further, participants who self-reported performing more web searches per day tended to have larger Google, eXelate, and BlueKai profiles, although the significance was weak ($p < 0.05$). Taken together, these findings resonate with our intuition that these three companies are more reliant on data from behavioral tracking than Facebook.

The most surprising results from our models concern privacy conscious behaviors. Almost none of these features are significant, including use of ad blocking extensions,[18] VPNs, or Private Mode browsing, as well as frequency of clearing browser history and cookies.

There are several possible explanations for this finding. *First*, participants could have misreported the use of privacy tools and privacy-enhancing behaviors. Many participants gave ambiguous, "I don't know" answers to some of these questions, suggesting a possible lack of familiarity with the area. *Second*, *APMs* could be tracking users through mechanisms that are resistant to the behaviors and tools we surveyed, such as browser fingerprinting [1], [40], [28], [27] or cross-device tracking [90]. *Third*, recent work from Bashir et al. [9] demonstrated that some anti-tracking browser extensions are not totally effective, meaning that participants using these tools may still be tracked. *Fourth*, *APMs*' relationships with data brokers may play a large role in the creation of interest profiles.

The only privacy feature that is strongly significant ($p < 0.001$) in our models is use of the DNT HTTP header. Across all four platforms DNT correlates with smaller interest profiles, although the effect is most significant for Google, and less so for Facebook. This finding remains consistent whether we treat "I don't know" responses from participants as "No" and encode the feature as binary (as we do in Table VI), or treat the feature as categorical. Facebook and Google do not honor DNT [83], so the cause of this correlation is unclear. One possibility is that the small number of participants who activated DNT (see Table V) engage in some other, related privacy-preserving behavior that we failed to measure in our survey.

V. Related Work

In this section, we first look at the studies documenting the pervasiveness of online tracking, tracking mechanisms used by advertisers, and user perceptions regarding online tracking. Then, we perform a brief survey of related work on *APMs*.

A. Online Tracking

Numerous studies have looked at the growth and privacy implications of web tracking [44], [41], [42], [43], [50], [27]. Similarly studies have documented the state of tracking on mobile devices [86], [11], [25], [33], [71], [70].

Researchers have closely examined the evolution of tracking mechanisms, including persistent cookies [37], local state through browser plugins [78], [5], browsing history through extensions [81], browser fingerprinting [57], [60], [72], [58], [1], [40], [28], and cross-device tracking [12], [90]. To maximize information about users, ad companies actively share tracking data with each other through cookie matching [1], [61], [29], [8], [9] and shared identifiers [1]. Through controlled experiments, Carrascosa et al. discovered that sensitive user attributes were being used to target online ads [14].

Researchers have proposed several techniques to mitigate online tracking, including: using machine learning to automatically identify trackers [51], [36], proposals for private cookies [63], and the addition of entropy into browsers to frustrate fingerprinting [56].

Uses of Tracking Data. Numerous studies have used controlled experiments to demonstrate that tracking data (and interests inferred from this data) does impact the ads shown to users [7], [52]. Additionally, there is evidence that some categories are more heavily targeted (e.g., insurance, travel, and real estate) [52], and that ads can be targeted to sensitive user attributes [89], [20], [88]. Interest profiles can also be used to target discriminatory ads [20], [80].

Perceptions of Tracking. Various surveys have found that people have concerns about the amount and type of information collected about them. McDonald et al. reported that 64% of the participants they surveyed found targeted advertising to be invasive [55]. Similarly, Turow et al. found that the majority of Americans feel that they do not have a meaningful choice with respect to the collection and use of their data by third-parties; thus, respondents were resigned to giving up their data [84]. Peoples' feelings about lack of agency may be rooted, in part, by widespread misconceptions about how targeted advertising systems are implemented [3], [54]. Balebako et al. discussed user concerns with respect to behavioral advertising and evaluated the effectiveness of privacy tools as counter mechanism [6].

Studies have found that a variety of factors influence people's perceptions of tracking and online advertising. Ur et al. reported that people found targeted advertising to be both useful and privacy invasive depending on how much they trusted the advertising company [85]. Similarly, Leon et al. surveyed 2,912 participants and found that they were willing to share information with advertisers if they were given more control over what was shared and with whom [49]. Like Ur et al., O'Donnell et al. surveyed 256 participants and found targeted advertising to be useful under a variety of circumstances (e.g., ads around major life events) [59]. However, Plane et al. found that people were very concerned when ad targeting resulted in discrimination (e.g., by targeting racial attributes) [67]. These findings highlight the complexity of peoples' relationship with target advertising, i.e., how trust, context, content, control, and effect commingle to shape perceptions of individual advertisements and the industry as a whole.

Dolin et al. surveyed people to understand how ad *explanations* (small disclosures near advertisements that provide insight into how the ad was targeted) impact peoples' opinions of targeted advertising. They found that peoples' comfort level varied based on the explanation they were given for how the targeted interest was inferred [24]. They also report that the accuracy of inferred interests was strongly, positively correlated with user comfort, regardless of the sensitivity of the

[18] Rather than include six binary features in our models (one for each extension, see Table V), we use a single feature that encodes the count of extensions that each participant reported having in their browser.

interest. Given that we observe that the majority of interests in *APM* profiles are not correct, this suggests that users would not be comfortable with ads that target these erroneous interests.

B. Ad Preference Managers

We are not the first to study *APMs*. Wills et al. [89] conducted controlled experiments to see the correlation between interests in Google's *APM* and the resulting ads that Google showed. They observed cases where interests were not visible in the *APM* despite observing corresponding behavioral ads. Datta et al. [20] also observed Google's *APM* omitting sensitive interests. Andreou et al. [4] observed missing attributes from Facebook's *APM*. These results are problematic, since *APMs* are supposed to allow users to see and adjust how they are being targeted for ads. These omissions also hinder studies that rely on *APMs* as a source of ground-truth data for experiments [34], [15], [7].

Degeling et al. used controlled experiments to examine how browsing different websites impacted the inferences drawn by BlueKai [22]. Interestingly, they found that identical browsing histories could produce different interest profiles, demonstrating that, at least for BlueKai, there is significant noise in the inference process. This observation may help explain our finding from § IV-C, that interests are often mis-inferred.

Google's *APM* used to show the gender and age that Google had inferred about users. To assess the accuracy of these inferences, Tschantz et al. performed a survey where they asked users to cut-and-paste this data from Google's *APM* and answer several questions [82]. They found that Google did not show these inferences 18–29% of the time, but when they did they were correct 65–74% of time. Like our study, the authors did not find significant correlations between the use of privacy-preserving tools/behaviors and the accuracy of Google's inferences.

To bring more transparency to this ecosystem, Lecuyer et al. [46], [47] built systems to infer the interests that Google constructs about users.

Unlike our work, **none** of these studies have compared interests from different *APMs*.

VI. Limitations

Our study is complex, and there are some limitations that are worth noting.

First, despite our best efforts, our participant sample is not representative of all web users: our participants skew young, technically sophisticated, and highly educated. Crowdsourcing workers are also more privacy-sensitive than users in general [38]. A larger, global, random sample of web users might reveal additional types of interests, although we suspect that our conclusions about the relative size, specificity, and overlap between *APMs* are likely to generalize.

Second, our study captures a snapshot of *APMs* at a specific point in time. Longitudinal data is needed to understand whether interest profiles from Google, Facebook, and eXelate grow indefinitely, or whether the platforms retire stale interests (Degeling et al. demonstrated that BlueKai does indeed replace old interests with fresh ones [22]). A related issue is that our historical data only covers several months of browsing and search activity. Although this data is fresh and should be highly relevant to advertisers, it is unknown whether older data would reveal the origins of additional interests.

Third, prior work has shown that *APMs* may hide sensitive interests (e.g., religion, alcohol, diseases, etc.) [89], [20], [4]. Our results should be interpreted as a lower bound on the quantity of interests collected by *APMs*. However, we hypothesize that *APMs* tend to show the majority of collected interests; otherwise users would trivially notice the omissions.

Fourth, with respect to understanding the origins of inferred interests, there are numerous potential data sources that are not covered by dataset, including browsing and search history older than 100 days, email tracking [26], activity from mobile apps, and activity from non-Chrome browsers.[19] Furthermore, we map each visited domain and search query into a single category; it is possible that our mapping does not correspond perfectly with the mapping used by advertisers, or that some activity may map into multiple categories. Lastly, it is possible that some interests are inferred using machine learning models that rely on emergent correlations between activity and seemingly unrelated categories.

VII. Concluding Discussion

In this work, we combine survey and crawled data to compare user interest profiles across four major advertising and tracking companies. For this study, we collected interest profiles and qualitative assessments of these interests from 220 participants. To facilitate comparison of interests across the *APMs*, we constructed and manually validated a shared term-space based on ODP categories.

Overall, we find that Facebook has the largest interest profiles of the four platforms we study (Figure 5), while BlueKai also has large, very specific interest profiles that are heavily predicated on information from data partners (Table II). We hypothesize that the four platforms rely on different underlying data sources, which leads to their respective profiles including drastically different interests for each given user (Figure 7).

Value and Policy. A key question raised by our study concerns the value of interest profile data. Participants in our sample only rated 27% of the interests in their profiles as strongly relevant (Figure 8), and this findings was consistent across *APMs*. Our results echo anecdotal findings from investigative journalists who have found their profiles from data brokers to be woefully inaccurate [48], [10]. Further, participants reported that ads targeted to these irrelevant interests were not useful (Figure 9). This raises difficult, unanswered questions about the online advertising industry: are the marginal utility gains of behaviorally-targeted ads relative to untargeted and contextually-targeted ads (if indeed there are gains) justified given their higher costs, in terms of money but moreso in terms of user privacy?

Concerns about online privacy have become so pervasive that countries are passing laws aimed at curtailing these practices (e.g., GDPR). And yet, despite the massive scale of tracking, we find that interest profiles remain imperfect.

[19] Chrome was used by >60% of web users worldwide circa-2018 [87]. Limiting our sample to Chrome users should not preclude obtaining a representative sample.

Industry may claim that this motivates greater and more invasive tracking, to achieve even more relevant user profiles. However, a privacy-centric interpretation is that the tracking industry may have passed the point of diminishing returns: no amount of additional data collection may ever result in "perfect" interest profiles. This suggests that it is worth investigating the tradeoffs between data minimization and efficacy of targeted ads, to determine the point in the space that yields greater privacy for users without substantially reducing the effectiveness of advertising campaigns or publishers' profits.

Provenance. Another key finding from our study is the extent to which interest profiles do not appear to be derived from participants' recent online behavioral data. We observe that participants' recent browsing and search history data explain less than half of the interests in our sample (Figure 10). This finding is especially interesting with respect to Facebook, which has a large web-tracking presence, and Google, which is the largest web tracker [27] and has logs of users' searches. As we discuss in § VI, there are a number of other possible data sources that may be the basis for the unexplained interests. Additionally, Degeling et al. note that non-determinism in tracking systems may also give rise to unexpected variations in inferred interests [22].

In our regression models, we find that privacy-conscious behaviors (e.g., clearing cookies) and privacy tools (e.g., ad blockers) have no statistically significant relationship with interest profile sizes (Table VI). One possible interpretation of this finding is that the behaviors and tools we surveyed are ineffective because browsing history is only one data source of many used to construct interest profiles. Alternative data sources include behavior in mobile apps [71], [70] or via email trackers [26]. Another potential explanation is that trackers have evolved to circumvent the countermeasures we surveyed. Indeed, ad networks are known to track users through other means like browser fingerprinting [1], [40], [28], [27], and prior work shows that blocking browser extensions are not entirely effective at their task [9].

One takeaway from these observations is that privacy-focused academics may need to shift their focus in the future. There is an enormous and important literature on web tracking, but we argue that the community must begin critically examining other information sources, especially data brokers. The same applies to privacy tools: preventing user data linkage across providers and devices may be as (or more) important today than blocking trackers for protecting individual privacy.

Acknowledgments

We thank the reviewers and our shepherd for their helpful feedback. This research was supported in part by NSF grants CNS-1563320 and IIS-1553088. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The web never forgets: Persistent tracking mechanisms in the wild," in *Proc. of CCS*, 2014.

[2] AdBlock, "Surf the web without annoying pop ups and ads!" get AdBlock., https://getadblock.com.

[3] L. Agarwal, N. Shrivastava, S. Jaiswal, and S. Panjwani, "Do not embarrass: Re-examining user concerns for online tracking and advertising," in *Proc. of the Workshop on Usable Security*, 2013.

[4] A. Andreou, G. Venkatadri, O. Goga, K. P. Gummadi, P. Loiseau, and A. Mislove, "Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations," in *Proc of NDSS*, 2018.

[5] M. Ayenson, D. J. Wambach, A. Soltani, N. Good, and C. J. Hoofnagle, "Flash cookies and privacy ii: Now with html5 and etag respawning," *Available at SSRN 1898390*, 2011.

[6] 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," in *Proc. of W2SP*, 2012.

[7] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and analyzing online display ads," in *Proc. of WWW*, 2014.

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

[9] M. A. Bashir and C. Wilson, "Diffusion of User Tracking Data in the Online Advertising Ecosystem," in *Proc. of PETS*, July 2018.

[10] T. Bergin, "How a data mining giant got me wrong," Reuters, Mar. 2018, https://www.reuters.com/article/us-data-privacy-acxiom-insight/how-a-data-mining-giant-got-me-wrong-idUSKBN1H513K.

[11] T. Book and D. S. Wallach, "A case of collusion: A study of the interface between ad libraries and their apps," in *Proc. of SPSM*, 2013.

[12] J. Brookman, P. Rouge, A. Alva, and C. Yeung, "Cross-device tracking: Measurement and disclosures," in *Proc. of PETS*, 2017.

[13] A. Cahn, S. Alfeld, P. Barford, and S. Muthukrishnan, "An empirical study of web cookies," in *Proc. of WWW*, 2016.

[14] J. M. Carrascosa, J. Mikians, R. Cuevas, V. Erramilli, and N. Laoutaris, "I always feel like somebody's watching me: Measuring online behavioural advertising," in *Proc. of ACM CoNEXT*, 2015.

[15] C. Castelluccia, M.-A. Kaafar, and M.-D. Tran, "Betrayed by your ads!: Reconstructing user profiles from targeted ads," in *Proc. of PETS*, 2012.

[16] L. Chang, "Yes, Facebook is reading the messages you send through Messenger," Digital Trends, Apr. 2018, https://www.digitaltrends.com/social-media/facebook-reads-messenger-messages/.

[17] G. Code, "word2vec: Tool for computing continuous distributed representations of words." Google., Jul. 2013, https://code.google.com/archive/p/word2vec/.

[18] N. Confessore, "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far," The New York Times, Apr. 2018, https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

[19] S. Cowley and J. Pepitone, "Google to pay record $22.5 million fine for Safari privacy evasion," CNNMoney, Aug. 2012, http://money.cnn.com/2012/08/09/technology/google-safari-settle/index.html.

[20] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination," in *Proc. of PETS*, 2015.

[21] DBpedia, "Towards a Public Data Infrastructure for a Large, Multilingual, Semantic Knowledge Graph," DBpedia., Jan. 2007, http://wiki.dbpedia.org.

[22] M. Degeling and J. Nierhoff, "Tracking and tricking a profiler: Automated measuring and influencing of bluekai's interest profiling," in *Proc. of WPES*, 2018.

[23] DMOZ, "About DMOZ," AOL Inc., Jun. 2016, http://dmoz-odp.org/docs/en/about.html.

[24] C. Dolin, B. Weinshel, S. Shan, C. M. Hahn, E. Choi, M. L. Mazurek, and B. Ur, "Unpacking Perceptions of Data-Driven Inferences Underlying Online Targeting and Personalization," in *Proc. of CHI*, 2018.

[25] M. Egele, C. Kruegel, E. Kirda, and G. Vigna, "Pios: Detecting privacy leaks in ios applications," in *Proc of NDSS*, 2011.

[26] S. Englehardt, J. Han, and A. Narayanan, "I never signed up for this! privacy implications of email tracking," *PoPETs*, vol. 2018, no. 1, pp. 109–126, 2018.

[27] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proc. of CCS*, 2016.

[28] S. Englehardt, D. Reisman, C. Eubank, P. Zimmerman, J. Mayer, A. Narayanan, and E. W. Felten, "Cookies that give you away: The surveillance implications of web tracking," in *Proc. of WWW*, 2015.

[29] M. Falahrastegar, H. Haddadi, S. Uhlig, and R. Mortier, "Tracking personal identifiers across the web," in *Proc. of PAM*, 2016.

[30] L. Faw, "Zenith: Google Remains Top-Ranked Media Company By Ad Revenue," MediaPost, May 2017, https://www.mediapost.com/publications/article/300315/zenith-google-remains-top-ranked-media-company-by.html.

[31] P. Gill, V. Erramilli, A. Chaintreau, B. Krishnamurthy, K. Papagiannaki, and P. Rodriguez, "Follow the money: Understanding economics of online aggregation and advertising," in *Proc. of IMC*, 2013.

[32] E. GmbH, "Adblock plus: Surf the web without annoying ads!" eyeo GmbH., https://adblockplus.org.

[33] M. C. Grace, W. Zhou, X. Jiang, and A.-R. Sadeghi, "Unsafe exposure analysis of mobile in-app advertisements," in *Proc. of WISEC*, 2012.

[34] S. Guha, B. Cheng, and P. Francis, "Challenges in measuring online advertising systems," in *Proc. of IMC*, 2010.

[35] A. Help, "How Google infers interest and demographic categories," Google, Jun. 2016, https://support.google.com/adsense/answer/140378?hl=en.

[36] M. Ikram, H. J. Asghar, M. A. Kâafar, B. Krishnamurthy, and A. Mahanti, "Towards seamless tracking-free web: Improved detection of trackers via one-class learning," *PoPETs*, vol. 2017, no. 1, pp. 79–99, 2017.

[37] S. Kamkar, "Evercookie - virtually irrevocable persistent cookies." September 2010, http://samy.pl/evercookie/.

[38] R. Kang, S. Brown, L. Dabbish, and S. Kiesler, "Privacy attitudes of mechanical turk workers and the u.s. public," in *Proc. of the Workshop on Usable Security*, 2014.

[39] R. Kattouw, V. Vasiliev, B. T. Minh, S. Reed, and Y. Astrakhan, "MediaWiki API," MediaWiki., Mar. 2008, http://thewikifire.org/api.php.

[40] T. Kohno, A. Broido, and K. Claffy, "Remote physical device fingerprinting," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 2, pp. 93–108, 2005.

[41] B. Krishnamurthy, D. Malandrino, and C. E. Wills, "Measuring privacy loss and the impact of privacy protection in web browsing," in *Proc. of the Workshop on Usable Security*, 2007.

[42] B. Krishnamurthy, K. Naryshkin, and C. Wills, "Privacy diffusion on the web: A longitudinal perspective," in *Proc. of WWW*, 2009.

[43] B. Krishnamurthy and C. Wills, "Privacy leakage vs. protection measures: the growing disconnect," in *Proc. of W2SP*, 2011.

[44] B. Krishnamurthy and C. E. Wills, "Generating a privacy footprint on the internet," in *Proc. of IMC*, 2006.

[45] D. Lachowicz, "Word Processing for Everyone," AbiWord., Apr. 2010, https://www.abisource.com/projects/enchant/.

[46] M. Lécuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu, "Xray: Enhancing the web's transparency with differential correlation," in *Proc. of USENIX Security Symposium*, 2014.

[47] M. Lecuyer, R. Spahn, Y. Spiliopolous, A. Chaintreau, R. Geambasu, and D. Hsu, "Sunlight: Fine-grained targeting detection at scale with statistical confidence," in *Proc. of CCS*, 2015.

[48] K. Leetaru, "The Data Brokers So Powerful Even Facebook Bought Their Data – But They Got Me Wildly Wrong," Reuters, Apr. 2018, https://www.forbes.com/sites/kalevleetaru/2018/04/05/the-data-brokers-so-powerful-even-facebook-bought-their-data-but-they-got-me-wildly-wrong/#68deb7183107.

[49] P. G. Leon, B. Ur, Y. Wang, M. Sleeper, R. Balebako, R. Shay, L. Bauer, M. Christodorescu, and L. F. Cranor, "What matters to users?: Factors that affect users' willingness to share information with online advertisers," in *Proc. of the Workshop on Usable Security*, 2013.

[50] A. Lerner, A. K. Simpson, T. Kohno, and F. Roesner, "Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016," in *Proc. of USENIX Security Symposium*, 2016.

[51] T.-C. Li, H. Hang, M. Faloutsos, and P. Efstathopoulos, "Trackadvisor: Taking back browsing privacy from third-party trackers," in *Proc. of PAM*, 2015.

[52] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan, "Adreveal: Improving transparency into online targeted advertising," in *Proc. of HotNets*, 2013.

[53] K. Liyakasa, "Nielsen Acquires Data Platform eXelate For Estimated $200 Million," Ad Exchanger, Mar. 2015, https://adexchanger.com/digital-tv/nielsen-acquires-data-marketing-company-exelate/.

[54] M. Malheiros, C. Jennett, S. Patel, S. Brostoff, and M. A. Sasse, "Too close for comfort: A study of the effectiveness and acceptability of rich-media personalized advertising," 2012.

[55] A. M. McDonald and L. F. Cranor, "Americans' attitudes about internet behavioral advertising practices," in *Proc. of WPES*, 2010.

[56] G. Merzdovnik, M. Huber, D. Buhov, N. Nikiforakis, S. Neuner, M. Schmiedecker, and E. R. Weippl, "Block me if you can: A large-scale study of tracker-blocking tools," in *IEEE European Symposium on Security and Privacy (Euro S&P)*, 2017.

[57] K. Mowery and H. Shacham, "Pixel perfect: Fingerprinting canvas in html5," in *Proc. of W2SP*, 2012.

[58] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "Cookieless monster: Exploring the ecosystem of web-based device fingerprinting," in *Proc. of IEEE Symposium on Security and Privacy*, 2013.

[59] K. O'Donnell and H. Cramer, "People's perceptions of personalized ads," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion, 2015.

[60] L. Olejnik, C. Castelluccia, and A. Janc, "Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns," in *Proc. of HotPETs*, 2012.

[61] L. Olejnik, T. Minh-Dung, and C. Castelluccia, "Selling off privacy at auction," in *Proc of NDSS*, 2014.

[62] Oracle, "Oracle Buys BlueKai," Oracle, Feb. 2014, https://www.oracle.com/corporate/acquisitions/bluekai/index.html.

[63] F. Papaodyssefs, C. Iordanou, J. Blackburn, N. Laoutaris, and K. Papagiannaki, "Web identity translator: Behavioral advertising and identity privacy with wit," in *Proc. of HotNets*, 2015.

[64] P. Parker, "Google Unveils Ad Preferences Center For Search And Gmail Ads," Search Engine Land, Oct. 2011, https://searchengineland.com/google-unveils-ad-preferences-center-for-search-and-gmail-ads-99292.

[65] T. Peterson, "FACEBOOK'S LIVERAIL EXITS THE AD SERVER BUSINESS," AdAge, Jan. 2016, http://adage.com/article/digital/facebook-s-liverail-exits-ad-server-business/302017/.

[66] PhantomJS, "Full web stack. No browser required," PhantomJS, Apr. 2011, http://phantomjs.org.

[67] A. C. Plane, E. M. Redmiles, M. L. Mazurek, and M. C. Tschantz, "Exploring user perceptions of discrimination in online targeted advertising," in *26th USENIX Security Symposium (USENIX Security 17)*, 2017.

[68] Prolific, "Bringing people together to power the world's research," Prolific., Apr. 2014, https://www.prolific.ac.

[69] L. RAINIE, S. KIESLER, R. KANG, and M. MADDEN, "Concerns About Personal Information Online," Pew Research Center, Sep. 2013, http://www.pewinternet.org/2013/09/05/part-2-concerns-about-personal-information-online/.

[70] A. Razaghpanah, R. Nithyanand, N. Vallina-Rodriguez, S. Sundaresan, M. Allman, C. Kreibich, and P. Gill, "Apps, trackers, privacy and regulators: A global study of the mobile tracking ecosystem," in *Proc of NDSS*, 2018.

[71] J. Ren, A. Rao, M. Lindorfer, A. Legout, and D. Choffnes, "Recon: Revealing and controlling pii leaks in mobile network traffic," in *Proc. of MobiSys*, 2016.

[72] F. Roesner, T. Kohno, and D. Wetherall, "Detecting and defending against third-party tracking on the web," in *Proc. of NSDI*, 2012.

[73] A. Senemar, "Facebook Partners With Shadowy 'Data Brokers' To Farm Your Information," Medium, Apr. 2016, https://medium.com/sherbit-news/facebook-partners-with-shadowy-data-brokers-to-farm-your-information-1129a5878b05.

[74] SimilarWeb, "Website Traffic Statistics & Market Intelligence," SimilarWeb LTD., Jan. 2013, https://www.similarweb.com.

[75] T. Simonite, "Facebook's Like Buttons Will Soon Track Your Web Browsing to Target Ads," MIT Technology Review, Sep. 2015, https://www.technologyreview.com/s/541351/facebooks-like-buttons-will-soon-track-your-web-browsing-to-target-ads/.

[76] M. Smith, "Android now the world's most popular operating system," CSO, Apr. 2017, https://www.csoonline.com/article/3187011/mobile-wireless/android-is-now-the-worlds-most-popular-operating-system.html.

[77] O. Solon, "Facebook can track your browsing even after you've logged out, judge says," The Guardian, Jul. 2017, https://www.theguardian.com/technology/2017/jul/03/facebook-track-browsing-history-california-lawsuit.

[78] A. Soltani, S. Canty, Q. Mayo, L. Thomas, and C. J. Hoofnagle, "Flash cookies and privacy." in *AAAI Spring Symposium: Intelligent Information Privacy Management*, 2010.

[79] G. O. Source, "An optical character recognition (OCR) engine," Google., Aug. 2006, https://opensource.google.com/projects/tesseract.

[80] T. Speicher, M. Ali, G. Venkatadri, F. N. Ribeiro, G. Arvanitakis, F. Benevenuto, K. P. Gummadi, P. Loiseau, and A. Mislove, "On the Potential for Discrimination in Online Targeted Advertising," in *Proc. of FAT**, 2018.

[81] O. Starov and N. Nikiforakis, "Extended tracking powers: Measuring the privacy diffusion enabled by browser extensions," in *Proc. of WWW*, 2017.

[82] M. C. Tschantz, S. Egelman, J. Choi, N. Weaver, and G. Friedland, "The accuracy of the demographic inferences shown on google's ad settings," in *Proc. of WPES*, 2018.

[83] L. Tung, "Google, facebook 'do not track' requests? fcc says they can keep ignoring them," ZDNet., Nov. 2015, https://www.zdnet.com/article/google-facebook-do-not-track-requests-fcc-says-they-can-keep-ignoring-them/.

[84] J. Turow, M. Hennessy, and N. Draper, "The tradeoff fallacy: How marketers are misrepresenting american consumers and opening them up to exploitation," Report from the Annenberg School for Communication, June 2015, https://www.asc.upenn.edu/sites/default/files/TradeoffFallacy_1.pdf.

[85] B. Ur, P. G. Leon, L. F. Cranor, R. Shay, and Y. Wang, "Smart, useful, scary, creepy: Perceptions of online behavioral advertising," in *Proc. of the Workshop on Usable Security*, 2012.

[86] N. Vallina-Rodriguez, J. Shah, A. Finamore, Y. Grunenberger, K. Papagiannaki, H. Haddadi, and J. Crowcroft, "Breaking for commercials: Characterizing mobile advertising," in *Proc. of IMC*, 2012.

[87] S. J. Vaughan-Nichols, "Chrome is the most popular web browser of all," ZDNet, Jan. 2017, http://www.zdnet.com/article/chrome-is-the-most-popular-web-browser-of-all/.

[88] G. Venkatadri, Y. Liu, A. Andreou, O. Goga, P. Loiseau, A. Mislove, and K. P. Gummadi, "Privacy Risks with Facebook's PII-based Targeting: Auditing a Data Broker's Advertising Interface," in *Proc. of IEEE Symposium on Security and Privacy*, 2018.

[89] C. E. Wills and C. Tatar, "Understanding what they do with what they know," in *Proc. of WPES*, 2012.

[90] S. Zimmeck, J. S. Li, H. Kim, S. M. Bellovin, and T. Jebara, "A privacy analysis of cross-device tracking," in *Proc. of USENIX Security Symposium*, 2017.