A Privacy Analysis of Cross-device Tracking

Sebastian Zimmeck, *Carnegie Mellon University;* Jie S. Li and Hyungtae Kim, *unaffiliated;*
Steven M. Bellovin and Tony Jebara, *Columbia University*

https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/zimmeck

This paper is included in the Proceedings of the
26th USENIX Security Symposium

August 16–18, 2017 • Vancouver, BC, Canada

ISBN 978-1-931971-40-9

A Privacy Analysis of Cross-device Tracking*

Sebastian Zimmeck, Jie S. Li, Hyungtae Kim, Steven M. Bellovin, Tony Jebara
School of Computer Science, Carnegie Mellon University
Department of Computer Science, Columbia University
szimmeck@andrew.cmu.edu, jl3620@nyu.edu, hk2561@columbia.edu,{smb, jebara}@cs.columbia.edu

Abstract

Online tracking is evolving from browser- and device-tracking to people-tracking. As users are increasingly accessing the Internet from multiple devices this new paradigm of tracking—in most cases for purposes of advertising—is aimed at crossing the boundary between a user's individual devices and browsers. It establishes a person-centric view of a user across devices and seeks to combine the input from various data sources into an individual and comprehensive user profile. By its very nature such cross-device tracking can principally reveal a complete picture of a person and, thus, become more privacy-invasive than the siloed tracking via HTTP cookies or other traditional and more limited tracking mechanisms. In this study we are exploring cross-device tracking techniques as well as their privacy implications.

Particularly, we demonstrate a method to detect the occurrence of cross-device tracking, and, based on a cross-device tracking dataset that we collected from 126 Internet users, we explore the prevalence of cross-device trackers on mobile and desktop devices. We show that the similarity of IP addresses and Internet history for a user's devices gives rise to a matching rate of F-1 = 0.91 for connecting a mobile to a desktop device in our dataset. This finding is especially noteworthy in light of the increase in learning power that cross-device companies may achieve by leveraging user data from more than one device. Given these privacy implications of cross-device tracking we also examine compliance with applicable self-regulation for 40 cross-device companies and find that some are not transparent about their practices.

1 Introduction

A recent study by Google showed that 98% of surveyed Internet users in the U.S. use multiple devices on a daily basis, and 90% switch devices sequentially to accomplish a task over time [37]. From an ad network's[1] perspective these developments create a challenging environment as they increase the complexity of targeting advertising to specific users. Attributing conversions of ads to actual purchases and frequency capping to avoid showing a user the same ad over and over again becomes more difficult as well. However, there is a solution to these challenges: cross-device tracking. This technique presents a fundamental shift from device tracking to people tracking. As shown in Figure 1, it principally allows ad networks to follow a user on his or her online journey through all devices. However, at the same time, cross-device tracking of users is potentially more privacy-invasive than the tracking of individual devices without connecting them.

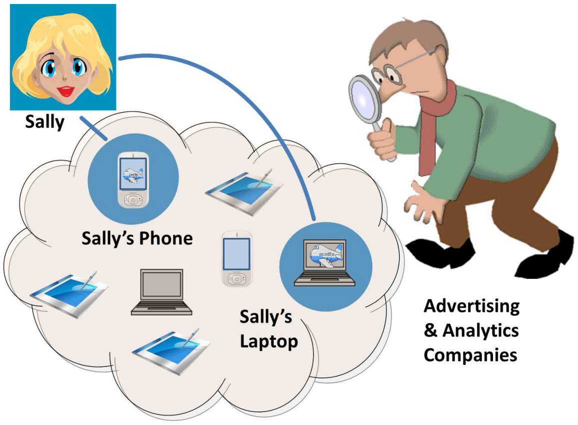


Figure 1: Identifying and correlating Sally's phone and desktop among all devices on the Internet allows cross-device companies to target ads on both of her devices.

In this study we are exploring the emerging cross-device tracking ecosystem from a privacy perspective. Particularly, we are interested in studying the tracking techniques used by cross-device companies,[2] understanding the extent to

[*] Most of the work on which we are reporting was done when Sebastian Zimmeck and Hyungtae Kim were at Columbia University.

[1] We are using the term ad network broadly encompassing ad exchanges, demand/supply side platforms, and other companies in the online ad space.

[2] The term cross-device company encompasses ad networks, analytics services, and other companies that are using cross-device tracking techniques.

which cross-device tracking occurs on desktop and mobile devices, and evaluating the privacy implications of machine learning applications to cross-device data. We understand cross-device tracking to mean the tracing of an individual's usage of the Internet on multiple devices and combining all resulting information into one comprehensive user profile.

Cross-device tracking exists in a deterministic and probabilistic variant. The former is based on a first-party relationship that often permits user identification with certainty, for example, when a user logs into a social network account from multiple devices. For the majority of our study we focus on probabilistic cross-device tracking, which is used by services that are limited to a third-party relationship with users. To that end, ad networks and analytics services oftentimes cooperate with web and app publishers that have a first-party user relationship and deploy tracking mechanisms on their properties. Applying machine learning they then correlate the various data streams to identify those belonging to the same users. Probabilistic and deterministic cross-device tracking approaches are often combined as companies of different provenance collaborate and exchange data [32]. While we examine cross-device tracking via HTTP cookies, pixel tags, and other traditional mechanisms, such tracking can also occur via ultrasound signals [7,31,59] or other side channels, which we do not examine here.

As some cross-device companies match billions of devices [22] and social networks have cross-device functionality naturally built into their systems lawmakers began to take notice. In particular, the U.S. Federal Trade Commission (FTC) hosted a cross-device workshop [29] facilitating an initial public discussion on the privacy implications of this form of Internet tracking. The regulators discussed with industry representatives, academics, and various other stakeholders privacy risks, consumer transparency, and effective industry self-regulation. They followed up with privacy recommendations for cross-device companies [32]. As evidenced by a recent case on cross-device tracking via ultrasound signals and the withdrawal of the service from the U.S. market, the FTC is determined to enforce the existing laws and regulations [31, 32], however, is hampered by insufficient insight into the used technologies [30].

In this study we are exploring cross-device tracking through the lens of privacy contributing the following:

1. By means of a brief case study we introduce a method for detecting cross-device trackers. We find statistical significance for various ad networks' capabilities of targeting mobile users on their desktop. (§ 3.)
2. We make publicly available a cross-device tracking dataset as well as software that we used for collecting the data.[3] We give a statistical overview of cross-device usage patterns for the users in our dataset. (§ 4.)
3. We design a basic algorithm and evaluate features and parameters for probabilistic cross-device tracking based on relevant patent and other industry documents. Using IP addresses, web domains, and app domains our techniques achieve an F-1 score of 0.91 on the collected data. (§ 5.)
4. Leveraging our dataset we analyze how the availability of both mobile and desktop data may impact the prediction of users' demographics and interests. Specifically, we examine predictions for gender and interest in finance. (§ 6.)
5. Based on our dataset we calculate the penetration of cross-device tracking on the Internet and conclude that some cross-device companies seem to have broad insight into Internet users' cross-device usage. (§ 7.)
6. Finally, we explore the efficacy of the industry's self-regulation and find that some cross-device companies do not transparently disclose their practices. (§ 8.)

2 Related Work

Our study is based on work in online tracking (§ 2.1), human-computer interaction (§ 2.2), and machine learning (§ 2.3).

2.1 Online Tracking

Much research was published on online tracking. Notably, Roesner et al. [72] developed a tracker taxonomy and examined how tracking occurs in the wild. Lerner et al. [56] provided a historical perspective of tracker evolution over time. However, few existing efforts discuss tracking *across* devices. Similar to traditional tracking such cross-device tracking requires the identification of individual users' browser instances. In this regard, Englehardt et al. [27] point out that cookies allow for linking a user's visits to different websites even if his or her device IP address varies. They conducted a large-scale measurement of traditional online tracking using their OpenWPM platform [26]. Cross-device tracking further requires the correlation of users' different devices. As Olejnik et al. [67] remarked, browsing histories could potentially identify the same user across multiple devices.

In the closest work to ours Brookman and his co-authors from the FTC [11] examine the potential for device correlation by surveying the occurrence of cross-device trackers on 100 popular websites. They also evaluate the extent to which cross-device companies notify users of their practices. While this inquiry into privacy transparency is part of our study as well (§ 8), we extend their work. In particular, we provide statistical support for the occurrence of cross-device tracking (§ 3), evaluate cross-device tracking techniques (§ 5), analyze the potential increase in learning power from cross-device data (§ 6), and examine the penetration of cross-device trackers (§ 7); all on real user data (§ 4). Our study is also complementary to the work of

[3] The dataset and software can be found at `https://github.com/SebastianZimmeck/Cross_Device_Tracking`.

Mavroudis et al. [59] and Arp et al. [7], who present analyses, attacks, and defenses for ultrasound-based cross-device tracking. We are exploring cross-device tracking based on cookies and other traditional tracking mechanisms.

If a browser does not accept cookies, it still can be tracked via device fingerprinting as initially shown by Kohno et al. [51], Eckersley et al. [25], and extensively surveyed by Lerner et al. [56]. Kurtz et al. [52] and Gulyás et al. [39] showed that mobile device fingerprints are often unique, distinguishable, and re-identifiable. Fingerprinting can be based on sensors [19] and, notably for our purposes, can be employed across browsers [15]. With their FPDetective Acar et al. [2] conducted a large-scale study of device fingerprinting. Nikiforakis et al. [66] provided insight into the practices of three popular browser-fingerprinting libraries and introduced PriVaricator [65], which is a defense against browser fingerprinting based on randomization. Three advanced tracking mechanisms—canvas fingerprinting, evercookies, and use of cookie syncing—were investigated by Acar et al. [1]. In our study we are now exploring the extent to which fingerprinting can play a role in cross-device tracking (§§ 5.1, 7.1). Various works on website fingerprinting [12,13,17,41,44,69] inform our study in this regard as well.

As we conduct a first cross-device tracking data flow experiment our work also relates to similar experiments and methodologies in other areas of online tracking. Particularly, our work relates to the study of Meng et al. [60], who showed that there is a correlation between Google ads and users' profiles and evaluated the likelihood of learning users' sensitive information. Focusing on Google as well, Lécuyer et al. [54, 55] were able to show a correlation between users' e-mail content and ads served to them. Further, Book and Wallach [10] collected a set of about 225K ads on 32 simulated devices and analyzed how the ads were targeted by correlating them to targeting profiles. In addition, Zarras et al. [85] performed a large-scale study on the security of ad serving, and Meng et al. [61] presented an ad fraud attack that enables publishers to increase their ad revenue. In our experiment we follow the recommendations given for information flow experiments by Tschantz et al [81].

2.2 Human-Computer Interaction

While there are only few online tracking studies investigating how users are tracked across devices, various efforts on human-computer interaction are informative for our purposes. The goal of these studies is to improve website navigation, browser prediction of user destinations, and search result relevance for search engines [3]. To that end, we leverage the insight of some studies focusing on website revisit patterns and highlighting the identifying potential of such revisits. In this regard, Tauscher and Greenberg [78] found that 58% of a user's visits to websites constitute revisits. People tend to access only a few pages frequently and browse in small clusters of related pages. Adar et al's [3] analysis reveals various patterns of revisits, each with unique behavioral, content, and structural characteristics.

Some studies took a closer look at website revisits across devices. Tossell et al. [79] were able to detect that revisits occurred very infrequently with approximately 25% of URLs revisited by each user. They further found that, compared to desktops, mobile browsers are accessed less frequently, for shorter durations, and to visit fewer pages. Users seem to rely on apps instead. Different from websites, apps have a revisit rate of 97.1% driven by a high number of visits to the five most frequently accessed apps. It appears that mobile web use is more concentrated and narrow than its desktop counterpart. Indeed, Kamvar et al.'s work [46] confirms this conjecture for the use of web search.

In their quest for improving the sharing of bookmarks and other information across devices Kane et al. [47] found that users tend to visit many of the same domains on both their mobile phones and desktops. Specifically, they found that a median of 75.4% of the domains viewed on the phone were also viewed on the desktop, and a median of 13.1% of the domains viewed on the desktop were also viewed on the phone. Despite the differing browsing habits across devices, particularly, the higher number of websites visited on desktops, they conclude that users' web browsing activities are similar across devices. However, users do not use all of their devices in the same way but rather assign them different roles, as Dearman and Pierce [20] found. As we will explore further (§ 6), these different roles could be a reason for why learning about users' interests can be more comprehensive with data from more than one device. While the sharing of devices can also principally impact cross-device companies' ability to track users across those, Matthew et al.'s study [58] found that phones were never shared among multiple individuals for mutual use, and computers were shared moderately.

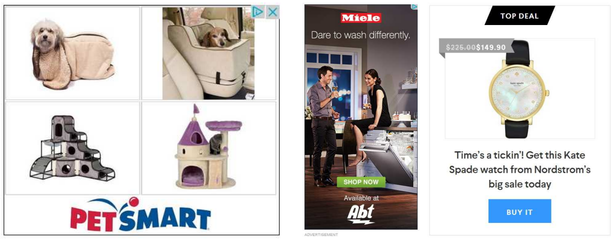
2.3 Machine Learning

Different from traditional types of online tracking cross-device tracking is often based on machine learning. In 2015, Drawbridge [22], an ad network specialized on cross-device tracking, hosted the ICDM 2015: Drawbridge Cross-Device Connections competition asking competition participants to leverage machine learning techniques to correlate devices to users [23]. The competition participants were given access to an anonymized proprietary dataset with mostly hidden features. The competition generated various short papers [14,48, 50,53,62,71,75,82] that take the perspective of an ad network. They focus on improving machine learning performance for a narrow set of features; essentially, only exploiting similarity of IP addresses. Our evaluation of tracking techniques broadens this research and is centered on privacy implications.

The first place solution of Walthers [82], which reached an F-0.5 score of 0.9, is in some ways representative for the

```
 1. google.com
 2. google.com; buy pet food - Google Search
 3. m.petsmart.com; PetSmart
 4. m.petsmart.com; Food
 5. m.petsmart.com; Fancy Feast Classic Adult Cat
 6. google.com; petco - Google Search
 7. m.petco.com; Pet Supplies, Pet Food, and Pet P.
 8. m.petco.com; Cat Furniture:  Cat Trees, Towers
 9. m.petco.com; Cat Food
10. m.petco.com; Browse & Buy Hill's Science Diet
11. m.petco.com; Hills Science Diet Adult Perfect W.
12. instinctpetfood.com; Instinct Pet Food
13. instinctpetfood.com; Instinct Pet Food For Your Cat
14. instinctpetfood.com; Instinct Raw for Cats
15. google.com; beneful cat food - Google Search
16. google.com; instacart
17. google.com
18. google.com; buy watch - Google Search
19. brilliantearth.com; Beyond Conflict Free Diamonds
20. google.com; buy refrigerator - Google Search
21. offers.geappliances.com; Drimmers - Offers GE A.
22. m.homedepot.com; Top Freezer Refrigerators - Re.
23. m.homedepot.com; Refrigerators
24. searshometownstores.com; Refrigerators & Freez.
25. searsoutlet.com; Refrigerators & Freezers for Sale
26. amazon.com
27. amazon.com; search for refrigerator
28. amazon.com; LG LSXS26366S 35-Inch Side
29. shoppermart.net; ShopperMart.net:  Find the best
30. samsung.com; Galaxy TabPro S - 2-in-1 Tablet
```

Figure 2: The mobile browser history (without visits to the Alexa-ranked homepages in the first two months of the experiment). The list shows the domains and the titles of the webpages, if any.


PETSMART
Miele
Dare to wash differently.
SHOP NOW
Available at
Abt
TOP DEAL
$225.00 $149.90
Time's a tickin'! Get this Kate Spade watch from Nordstrom's big sale today
BUY IT

A. PetSmart
nytimes.com
adsense.com
Google AdSense
Google Display Network

B. Miele/Abt
latimes.com
as.chango.com
Rubicon Project
Tapad

C. Kate Spade
aol.com
redirectingat.com
Skimlinks
Lotame

Figure 3: Selected ads served to the desktop browser after visiting the sites in Figure 2 on the mobile browser. We had not seen any of these ads in our desktop browser session two months before.

approaches taken in the competition. Comparable to other participants' solutions [14, 50, 53], it identified IP addresses that devices of the same user were connected to as the most important feature. Intuitively, as conjectured by Cao et al. [14], devices with similar IP footprints are more likely to be used by the same individual. Thus, the simple reliance on IP address history can already lead to an F-0.5 score of 0.86 [14]. However, various studies found that not all IP addresses are equally meaningful, in particular, because the same public or cellular IP address can be assigned to many different users at different times [48, 82].

Participants in the Drawbridge competition [23] did not find online history particularly useful for their task. They reported that correlating online history across devices provided only minimal gain [50, 53]. This seemingly contradictory result to the previously discussed usability studies, which hinted at cross-device website revisit patterns as an important feature, could be due to the fact that the Drawbridge dataset [45] provided only app history for mobile devices. Thus, the absence of mobile web history could be a reason for participants' inability to reach Drawbridge's precision of 0.97 [22]. While the Drawbridge competition was about the correlation of different user devices, it did not address the purpose of the correlation: the prediction of users' demographics and interests, which we will discuss in our study (§ 6).

3 Detecting Cross-device Trackers

In order to evaluate the occurrence of cross-device tracking in the wild we conducted an exploratory case study.

Purpose. It could be argued that our case study is aimed at the obvious: detecting the existence of cross-device tracking. However, we emphasize that it is our intention to show a procedure for identifying *unknown* cross-device companies. The procedure is also intended to be used for determining whether known companies are adhering to the limits that (self-)regulation imposes on them, in particular, as users are given the right to opt out from cross-device tracking [21]. Our case study provides an initial information flow experiment in the cross-device space. However, we caution that we leave a comprehensive analysis, which was done in other areas of online tracking [54, 55], for further research.

Establishing an IP Address Connection. We began our experiment by connecting two devices—a desktop and a mobile device—to the same router and modem. Using a fresh desktop browser without any user data we visited the homepages of five randomly selected news websites from the Alexa rankings [4]—aol.com, latimes.com, nytimes.com, wsj.com, and washingtonpost.com (the test homepages). We refreshed each test homepage ten times, as recommended for these type of information flow experiments [81], and observed the ads that were served. We also set up a desktop device with a fresh browser connected to a different router and modem as control instance. In the following two months we occasionally and randomly visited 100 highly ranked homepages [4] on our fresh mobile browser.

Observing Cross-device Ads. After two months we used the mobile browser to visit the websites shown in Figure 2. We searched Google for various consumer products and clicked on ads served for those on the search results pages. After a few hours we switched to our desktop browser and accessed the test homepages. We refreshed each ten times. Some of the served ads, which we had not seen before, were for products we had searched for on the mobile device. Figure 3 shows the ads and associated information, that is, the name of the ad (e.g., PetSmart), the domain on which it was served (e.g., nytimes.com), the domain of the ad server (e.g., adsense.com), the ad network serving the ad (e.g., Google AdSense), and the involved cross-device tracking provider (e.g., Google Display Network).[4] Our results suggest that the ad networks serving the ads had learned that the user who did the search on the phone was the same as the user on the desktop.

To assess the similarity of ads we categorized each ad according to Google ad categories [35]. Then, based on an exact one-tailed permutation test, as recommended [81], we compared the ad distribution served on the desktop browser to the ad distribution served on the desktop control browser. We evaluated the null hypothesis that both distributions do not differ from each other at the 0.05 significance level. However, the result of $p = 0.02$ indicates that the null hypothesis should be rejected and that the deviation of both distributions is statistically significant at the 0.05 level. This finding suggests that we successfully identified instances of cross-device tracking. We also found mobile cookie syncing between Rubicon Project and Tapad. However, confirming earlier observations [11], we did not detect any cookie syncing across devices.

Direction of Ad Serving and App-Web Correlation. In addition to cross-device tracking from mobile to desktop we were further interested in the reverse direction. However, searching Google on our desktop for buying products did not seem to lead to ads for these products on our mobile browser. One explanation might be that the ad serving was limited to one direction—from mobile to desktop— as users tend to move from a smaller to a larger screen [33, 37]. Another explanation could be that ad networks attached more weight to the history on the device to which an ad was served and less to other connected devices. Further, we might simply have missed all cross-device campaigns at the time for the products we searched for. Finally, we were not able to notice any correlation in ad serving in either direction when repeating our experiment with mobile apps instead of websites.

4 The Cross-device Tracking Dataset

A major reason for the scarcity of academic research in cross-device tracking is the unavailability of data. Generally, only proprietary industry data exists.[5] Thus, we collected our own cross-device tracking dataset (the CDT dataset). Here we describe how we collected the data and highlight cross-device usage patterns of the users in the dataset.

	Desktop Web	***Mobile Web***	***Mobile Apps***
Users	125	102	104
IPs	**1,994**	**5,784**	
Domains	**23,517**	**3,876**	**845**

Table 1: Summary statistics showing the total number of unique users, IP addresses, and domains in the CDT dataset.

	Desktop Web	***Mobile Web***	***Mobile Apps***
	25th, 50th, 75th	*25th, 50th, 75th*	*25th, 50th, 75th*
Days	19, **22**, 26	9, **17**, 23	19, **22**, 24
IPs	6, **17**, 24	25, **63**, 92	
Domains	149, **251**, 374	9, **31**, 70	19, **30**, 44

Table 2: Summary statistics for the CDT dataset per user showing the unique values at the 25th, 50th, 75th percentiles. The data was collected for the same continuous time period for every user. However, not every user made use of his or her devices every day.

Data Collection Procedure. Before we began the data collection we obtained approval from Columbia University's Institutional Review Board (IRB). We built our data collection system such that interested users could sign up on our project website, at which point we also took a device fingerprint for each signed up device. We asked users to supply basic information on their demographics (e.g., age and gender), interests (e.g., finance, games, shopping) [36], and personas (e.g., avid runners, bookworms, pet owners) [84]. In order to capture users' mobile and desktop history we provided them with browser extensions and an Android app that we developed for automatically collecting such information.[6] Details on the types of information that we collected are contained in Appendix A. We do not have any indication that users behaved differently in our study than under real-world conditions.

We only signed up users of Android phones with Android's native browser, Google Chrome, or the Samsung S-Browser. We did not support iOS or other operating systems. Our app requires Android 4.0.3 and runs without root access. Every minute it checks for a new foreground app running on the device as well as new entries in the browsing history database of the phones' browsers. If new apps or URLs are detected, a new history datapoint is transmitted to our server.[7] On the desktop side we provided users of all operating systems with data collection browser extensions for Google Chrome, Mozilla Firefox, and Opera. At the conclusion of the study we rewarded each user with an Amazon gift card for $15 to $50 depending on the amount of data we received from them.

Dataset Characteristics. We collected data from 126 users. Tables 1 and 2 show further details. We signed up 125 desktop and 108 mobile users with an intersection of 107

[4] We assume that the Google Display Network covers sites using AdSense, DoubleClick, Blogger, YouTube, and AdMob. On one side, this is likely an overestimation as not all sites using these trackers are part of the Google Display Network. On the other side, it is an underestimation as there are sites that are part of it, however, not using any of the trackers. In total, the Google Display Network covers over two million sites [34].

[5] The Drawbridge dataset [45] was only accessible to participants of the Drawbridge competition [23] and limited in its use for that purpose.

[6] When we refer to desktops, we include laptops but exclude tablets.

[7] For some users with Google Chrome and Android 6.0 or higher we did not receive the full browsing history due to browser restrictions. We asked affected users to send us their history manually.

users from whom we obtained both mobile and desktop data. While our data faithfully represents that not every Internet user has multiple devices, it does not reflect that users in the real world can have more than two devices. However, despite this limitation we believe that our dataset is generally an accurate reflection of real multi-device usage on the Internet because the vast majority of mobile devices is associated with only one desktop browser [71]. Therefore, it seems plausible to adopt this understanding of the problem here as well. Further, only 3/108 (3%) of mobile users and 4/125 (3%) of desktop users in our study reported that they are sharing their devices. As this result seems in line with findings that phones are never shared for mutual use and that computers are only shared for a moderate amount [58], it appears that our data is a realistic representation in this regard.

118 users in our study were affiliates of Columbia University, mostly students. Based on this population we believe that our data is more homogeneous than a data from, say, the general population of New York City. However, we also note that our users are less likely to encounter typical restrictions of device use that many employees face in the workplace, e.g., corporate networks blocking certain websites. For the median user we collected about three weeks of data of which IP addresses and domains are of particular importance for probabilistic cross-device tracking because they can be used to measure the similarity between devices (§ 5.2).

It is noteworthy that the total unique mobile IP count (5,784) is nearly three times the total unique desktop IP count (1,994), which reflects mobile usage on the go. It should be noted, though, that the real unique mobile IP count is likely even higher as our method did not allow us to collect mobile IPs with every datapoint. However, the high number of unique desktop domains (23,517), compared to the homogeneous usage of apps (845), underscores the diversity of desktop browsing. While it is much more diverse in terms of domains (3,876), mobile web usage pales in comparison to app usage. As shown by the 25th, 50th, and 75th percentiles, the median user accessed the mobile web only for 17 days visiting only 31 unique domains.[8] While app usage is more popular with a median of 22 days, the median usage of 30 unique apps is comparable to that of the mobile web. However, the median number of unique mobile IP addresses (63) more than triples desktop IP addresses (17).

Figure 4 shows that many users visit a relatively large number of unique mobile device IP addresses and desktop web domains. However, there does not seem to be a correlation between desktop and mobile devices to the effect that lower usage of one would imply more usage of the other or that both are used to an equal degree.

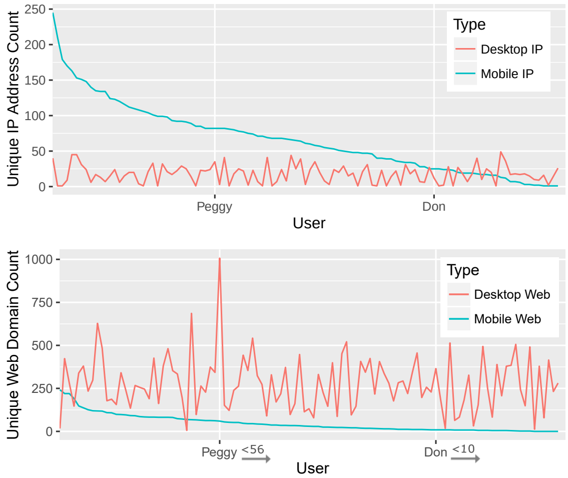


Figure 4: Unique IP address (top) and web domain (bottom) count for each user in our dataset for whom we had both mobile and desktop data. For example, Peggy has 82 unique mobile and 35 unique desktop IP addresses (top). To the right of Peggy about two thirds of users visited fewer than 56 unique mobile domains and to the right of Don about a fourth visited fewer than ten (bottom).

5 Methods for Cross-device Tracking

How cross-device companies operate is not known in detail [49]. In order to get an understanding of their capabilities we designed an algorithm and evaluated features and parameters informed by a review of public materials, particularly, Adelphic's cross-device patent [83] and Tapad's patent application for managing associations between device identifiers [80]. Essentially, cross-device tracking is based on resolving two tasks: first, uniquely identifying users' devices (§ 5.1), and, second, correlating those that belong to the same user (§ 5.2).

5.1 Identifying Devices

Traditionally, HTTP cookies are used to identify desktop devices. Indeed, many cross-device companies are employing cookies for their tracking purposes as well. For mobile devices the use of advertising identifiers, such as Google's Advertising ID (AdID), is common and often combined with cookie tracking. Thus, if users are allowing cookies and do not opt out from being tracked, both their mobile and desktop devices can be easily identified. However, with the surge of tracking- and ad-blocking software, which some consider a mainstream technology on mobile by now [68], unconventional identification technologies, such as device fingerprinting, are becoming more prevalent. While it does not appear that they will generally replace cookies and advertising identifiers any time soon, various cross-device companies—for example, BlueCava [9] and AdTruth [28]—are making use of device fingerprinting.

[8] A day was counted if a user's device had at least one desktop web, mobile web, or app access on a given day. Also, uniqueness of a domain is dependent on its top and second level. Thus, for example, we treat facebook.com and linkedin.com as different domains, however, linkedin.com and blog.linkedin.com as the same domain.

	Desk Devices $H, H_n, \hat{H}$	***Mob Devices*** $H, H_n, \hat{H}$
User Agent	**4.46**, 0.64, 4.96	**6.43**, 0.95, 8.5
Display Size/Colors	5.34, 0.77, 6.08	1.72, 0.25, 2.08
Fonts	6.11, 0.88, 7.33	1.21, 0.18, 1.33
Accept Headers	2.86, 0.41, 3.29	2.34, 0.35, 3
System Language	0.41, 0.06, 0.51	0.81, 0.12, 1
Time Zone	0.25, 0.04, 0.35	0.53, 0.07, 0.74
Mobile Carrier	N/A	1.39, 0.21, 1.45
Do Not Track Enabled	0.67, 0.1, 0.67	0.18, 0.03, 0.19
Geolocation Enabled	0.45, 0.07, 0.45	1, 0.15, 1
Touch Enabled	0.72, 0.1, 0.72	N/A
Total per Device Type	**6.96**, **1**, **12.95**	**6.69**, **0.99**, **10.87**
Total	**7.84**, **1**, **13.37**	

Table 3: Entropy (H), normalized entropy (H_n), and estimated entropy ($\hat{H}$) for various browser features in our CDT dataset. The normalized entropy ranges from 0 (all features are the same) to 1 (all features are different). We calculated the estimated entropy according to Chao and Shen [16]. For the totals we considered all listed features. Overall, our dataset contains 3 duplicate mobile fingerprints and 1 duplicate desktop fingerprint.

Cross-device companies that are solely relying on device fingerprinting must be able to identify both desktop and mobile devices using this technique. While it was reported that device fingerprints generally do not work well on mobile devices [25], our results do not support such broad conclusion. Particularly, mobile user agents often contain distinctive features and are far more diverse (6.43 bits) than user agents on desktops (4.46 bits). Also, the entropy in our dataset only represents a lower bound as we imposed substantial limitations for users' participation in our study; most notably, requiring them to have an Android phone with Android 4.0.3 or higher and use the native browser, Chrome, or S-Browser. We also did not consider, for instance, canvas fingerprinting [1], sensor data [19], or the order in which fonts and plugins were detected [25]. However, most mobile devices in our dataset were still identifiable. The detailed findings for the 107 mobile and 126 desktop devices in our CDT dataset are shown in Table 3.[9] Due to the small size of our dataset we caution to interpret our results as indicative for the reliability of mobile device fingerprinting, though.

5.2 Correlating Devices

After uniquely identifying each device cross-device companies must match those that appear similar. Successfully matching devices at scale is the core challenge for cross-device companies. Devices are represented in graphs known as Device Graphs [76], Connected Consumer Graphs [22], or under similar proprietary monikers. From a graph-theoretical perspective a device graph can be built from connected components (each of which represents a user) with a maximum number of vertices (devices) and edges (device connections) [18]. Matching every mobile with exactly one desktop device will result in a bipartite graph. The goal is to achieve a perfect matching of similar devices.

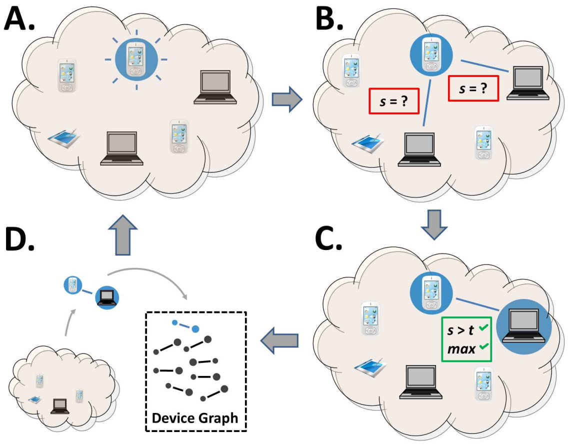


Figure 5: Our cross-device tracking approach. A. First, a mobile device is identified. B. Its similarity to each identified desktop device, s, is calculated. C. The mobile-desktop pair with the maximum similarity, *max*, that is above a similarity threshold, t, is determined, if any. D. If such pair exists, it is added to the device graph and the next iteration starts with a new mobile device. This routine is repeated in three consecutive stages each evaluating similarities between mobile and desktop IP addresses, mobile and desktop URLs, and mobile apps and desktop URLs, respectively. If a mobile device cannot be matched in one stage due to not overcoming the similarity threshold, a match is attempted in the next.

Algorithm, Features, and Parameters. While deterministic cross-device companies can simply match a user's devices based on his or her login information, which may also extend towards third party properties through single sign-on functionality, achieving a high match rate is more difficult for probabilistic cross-device tracking companies. In the Drawbridge competition [23] many participants applied gradient boosting [48, 50, 53, 62, 71]. However, some participants also combined support vector machines and factorization models into field-aware factorization machines [75] or employed pairwise ranking and ensemble learning techniques [14]. Interestingly, the best performing solution relied on learning-to-rank models instead of using the more conventional binary classification models [82].

In our approach, as outlined in Figure 5, we determine the similarity between devices based on distance metrics, most notably, the Bhattacharyya coefficient, which is defined for the distributions p and q as $Bhatta(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)}$. The use of distance metrics for device correlation was described in Adelphic's cross-device patent [83]. Our cross-device tracking algorithm works in multiple stages. Using a key insight from the patent, for each feature a

[9] One user did not submit a mobile fingerprint and another user submitted two different desktop fingerprints.

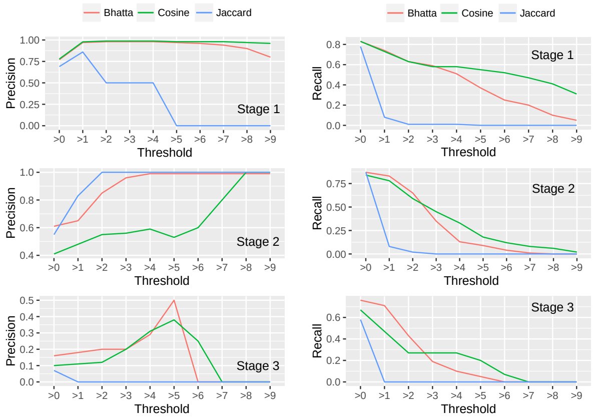
	Sim Feature Mapping	*Distance Metric*	*Sim Thresh*	*Mean Sim*	n_q	*Acc*	*Prec*	*Rec*	*F-1*
Stage 1	Mob IPs to Desk IPs	Bhatta	0.07	0.33	44	0.61	1	0.63	0.77
Stage 2	Mob URLs to Desk URLs	Bhatta'	0.13	0.18	44	0.52	0.85	0.59	0.7
Stage 3	Mob Apps to Desk URLs	Bhatta*	0.02	0.11	44	0.16	0.19	0.5	0.27
Stages 1–3	Same as in Individual Stages above			0.33, 0.16, 0.03	44	**0.84**	**0.88**	**0.95**	**0.91**

Table 4: Test set results. The first three rows show the results for running each stage individually. The fourth row shows the results for running the three stages consecutively. We normalized the Bhattacharyya coefficient (Bhatta) to a range between 0 (low similarity) and 1 (high similarity). Bhatta' denotes the exclusion of URLs in the Alexa Top 50 [4] and all columbia.edu URLs. Further, Bhatta* excludes the most used 100 apps according to our training set. We selected the best similarity threshold (Sim Thresh) for each stage according to observations in our training runs. Mean Sim is the mean similarity across all 44 device pairs in the test set, and n_q is the size of the test set.

similarity threshold is set. If a threshold is reached at one stage, a match is declared for the mobile-desktop device pair with the highest similarity score. Otherwise, the algorithm continues to evaluate whether the similarity threshold for a different feature is reached in the next stage. To evaluate the similarity between a mobile and a desktop device it compares mobile to desktop IP addresses, mobile to desktop web URLs, and mobile apps to desktop web URLs.

Test Set Results. To test our approach we randomly separated the set of device pairs in our dataset into a training ($n_t =$ 63) and a test set ($n_q = 44$). We used the former to tune our algorithm and features and held out the latter for performance evaluation. As shown in Table 4, running all three stages of the algorithm consecutively on our test set leads to precision, recall, and F-1 scores of 0.88, 0.95, and 0.91, respectively. The F-0.5 score [50], which emphasizes precision over recall, reaches 0.91. In detail, we obtained 37 true positives (TP), 5 false positives (FP), 0 true negatives (TN), and 2 false negatives (FN). These results are based on the usual definitions, i.e., accuracy, $Acc = (TP+TN)/(TP+TN+FP+FN)$, precision, $Prec = TP/(TP+FP)$, recall, $Rec = TP/(TP+FN)$, and F-1 score, $F\text{-}1 = (2 \cdot Prec \cdot Rec)/(Prec+Rec)$.

To make the matching more difficult we included in each run of our algorithm in every stage data from users for which we only had data from one device type: data from one user who only submitted mobile data and from 18 users who only submitted desktop data. Further, our results are based on modeling device correlation as binary classification. Specifically, for each correct match between a user's mobile and desktop device we counted a true positive. For each incorrect match we noted a false positive. If a mobile device would have no corresponding desktop device it would have been counted as a true negative if it remained unmatched. However, as there was only one such instance in our test set and that mobile device was actually matched, we counted it as false positive. A false negative means that an instance should have been matched, however, remained unmatched.



	Bhatta *F-1, Sim Thresh*	***Cosine*** *F-1, Sim Thresh*	***Jaccard*** *F-1, Sim Thresh*
Stage 1	**0.84**, 0.1	0.83, 0.1	0.73, 0
Stage 2	**0.74**, 0.2	0.59, 0.1	0.67, 0
Stage 3	**0.29**, 0.1	0.29, 0.4	0.12, 0

Figure 6: Precision and recall for matching devices based on various distance metrics and thresholds. The table shows the best F-1 scores and their corresponding similarity thresholds. The features are the same as described in the respective stages in Table 4. However, the evaluation is performed here on the full dataset. For higher thresholds recall scores tend to decrease while precision scores tend to increase (except when they exclude too many true positives). Overall, the Bhattacharyya coefficient returns the best results.

Running the three stages of our algorithm consecutively leads to approximately balanced results for precision (0.88) and recall (0.95), as shown in Table 4. However, when running the stages individually, we obtain relatively higher precision and lower recall in the first two stages and lower precision and higher recall in the third stage. This difference highlights the tradeoff between achieving correct matches (precision) and broad user coverage (recall). While it is challenging to improve one without adversely affecting the other [49, 73], the similarity thresholds provide the controls for adjustment. Figure 6 shows changes in precision and recall for different similarity thresholds and distance metrics.

The high precision scores of Drawbridge (0.97 [22]) and Tapad (0.91 [77]) seem to suggest that the industry favors precision over recall.[10] However, there is also an argument

[10] We interpret Tapad's usage of the term accuracy to mean precision ("[W]henever our Device Graph indicated a relationship between two or more devices, it was accurate 91.2 percent of the time.").

to be made against emphasizing precision: some device mismatches may be irrelevant. Particularly, we believe that mismatches might happen for people living in the same household (in case of mobile IP to desktop IP similarity) or individuals having the same interests (in case of web domain and app to web domain similarity). In these situations a mismatched device might still be a meaningful ad target [24]. The reason is that targeted purchase decisions might be made at the household level or look-alike audiences might be sufficiently valuable for an ad network [80].

Our results show that IP addresses are very meaningful for matching devices, which is in line with Cao et al.'s findings [14]. They reached an average F-0.5 score of 0.86 in the Drawbridge competition [23] using only features from IP address data. However, beyond this finding our results further suggest that visited web domains are a good indicator for device similarity as well. In fact, there might be situations in which they can be more revealing than IP addresses. For example, if users of the same household share an IP address, their devices can not be distinguished based on this feature. Also, while the correlation between apps and desktop domains does not contribute as much as the IP address and domain correlations, it still provides some meaningful signal as the results for the individual run of the third stage in Table 4 demonstrate. Most importantly, however, performance seems to increase if multiple features are applied consecutively. Some users can be better matched based on IP addresses and others on web domains or apps.

We note that we leveraged a manual mapping between apps and desktop domains via company names or other common identifiers thereby transforming a feature with minimal effect in our dataset and the Drawbridge competition [14,48,50,53,62,71,75,82] to a useful feature. Similarly, the domain mapping proved to be useful as well due to users' visits to the same domains across devices. These results highlight that cross-device matching is not completely reliant on IP matching, as suggested by the results in the Drawbridge competition. Our results seem to confirm the conjecture that carefully hand-crafted similarity features are of paramount importance while algorithms play a smaller role for the task of correlating mobile and desktop devices [82].

We experimented with various other features that ultimately did not prove useful. In particular, an algorithm leveraging system language and time zone did not match devices better than random. We also tried excluding sets of frequently used public IP addresses. However, different from excluding domains and apps, which, as described in Table 4, proved to be beneficial, this measure did not lead to better performance. We further tried different matching thresholds and evaluated various distance metrics as shown in Figure 6. In future work it would be interesting to examine the extent to which the time, order, and duration of app and url access play a role for device correlation. E-mail and other message content is an obvious candidate for a useful feature as well.

Applicability to Larger Datasets. With a runtime of $\mathcal{O}(n(n-1)/2)$ our algorithm is suitable for large scale analysis. However, it is obvious that our dataset is many orders smaller than the data that cross-device companies are usually working with. This difference in size begs the question to which extent our findings are applicable to larger datasets. For the similarity of IP addresses this question was already reliably answered. The Drawbridge competition results, for instance, by Landry et al. [53], are based on a set of about 62K mobile devices and confirm the meaningfulness of IP features. For web domain features the situation is different as the Drawbridge data did not contain those for mobile devices. However, we can make an argument that lends some supports for the applicability of our results to larger datasets.

Whether web data can be correlated across devices rests on two premises: first, users visiting an intersecting set of domains on both their mobile and desktop devices and, second, domains being sufficiently distinct to allow identification of users. To examine the first premise we randomly selected 50 U.S. domains out of the top 5K sites that were quantified by Quantcast [70] and found a mean of 17.1% users visiting a website both on a mobile and desktop device (during a 30-day period and at the 95% confidence level with a lower bound of 14.4% and an upper bound of 19.5% using the bootstrap technique). As to the second premise, it was shown for a set of about 368K desktop and mobile Internet users that 97% of them were uniquely identifiable if at least four visited websites were known.[11]

Limitations. It would be an interesting exercise to compare our techniques against those currently in use in industry. However, we are not aware of any publicly available resources allowing us to do so. The same is true for cross-device tracking datasets. To our knowledge, there is no dataset publicly available beyond the CDT dataset that we created. The only other cross-device tracking dataset we know of was made available by Drawbridge to participants of the Drawbridge competition solely for competition purposes [23]. However, even if this dataset would be available, it would only allow an incomplete analysis, particularly, as features were generally anonymized and mobile web history was not included in the dataset. Consequently, at this point it does not seem possible to compare our approach to others or evaluate its performance on a different dataset. However, as we implemented key design elements that we found in available industry materials, we believe that our results provide a first approximation for cross-device tracking approaches applied in practice.

There are various considerations of identifying and correlating devices in practice that we cannot meaningfully test. A first point concerns the time period for which users are being tracked. We believe that the three weeks of data that we have available for most users (Table 2)—the concrete

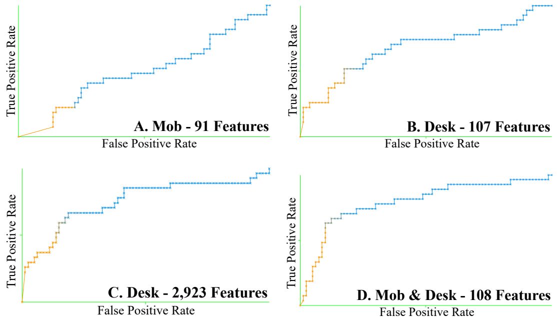
[11] All users in our dataset who visited at least one mobile and one desktop website had unique web histories as well.

length depending on the number of days on which they used their devices—are realistic. However, we lack the insight for which duration cross-device tracking actually occurs in practice. Also, despite some cross-device companies' broad coverage of websites and apps (§ 7), none of them has access to complete IP, web, and app data of Internet users. However, ultimately this limitation is one of reach and not of performance. By setting similarity thresholds high even companies with limited data can obtain precise results, albeit, at the cost of low recall. Further, our dataset does not contain full IP histories either. In addition, our data is probably more homogenous than real data, and, thus, more difficult to assess. Users in our study were mostly students located in a confined space with many commonly shared web domains and IP addresses.

6 Learning from Cross-device Data

In this section we examine whether cross-device data enables cross-device companies to make more accurate predictions than they could make using data from individual devices alone. We address this question for two inquiries: users' interest in finance—a randomly selected interest category—and gender. Both are relevant ad targeting criteria. For interest in finance we obtained the most accurate predictions by using data from both mobile and desktop devices. Consequently, this is a task in which predictions about a user from cross-device data appear more privacy-invasive than those from single device data sources. However, as we also found a lack of performance increase for predicting a user's gender, it appears that some prediction tasks might not become more accurate with the availability of cross-device data.

Predicting Interest in Finance. As a starting point for our feature creation we used Alexa category rankings [5] and Google Play store categories [38] to identify the top 25 finance domains that have both a website and an app. Then, we used the Weka machine learning toolkit [40] to explore the potential for predicting interest in finance. We experimented with various features and all available standard algorithms. We used a word-to-vector preprocessor and found logistic regression to be the most effective technique. Due to the class imbalance of only 23% users in our dataset expressing an interest in finance we ran logistic regression as a cost-sensitive classifier increasing the cost for a false positive on average 1.5 times over the cost for a false negative. Our results, which are shown in Figure 7 and based on 10-fold cross validation, suggest that predicting an interest in finance for users in our dataset is more accurate if both desktop and mobile data are available.



	Acc	***95% CI***	***Prec***	***Rec***	***F-1***	***ROC***
A.	0.64	0.55–0.73	0.26	0.22	**0.24**	**0.5**
B.	0.75	0.67–0.83	0.5	0.52	**0.51**	**0.68**
C.	0.79	0.71–0.87	0.57	0.59	**0.58**	**0.75**
D.	0.83	0.76–0.9	0.68	0.63	**0.65**	**0.79**

Figure 7: Logistic regression for predicting interest in finance from mobile web domains and apps (Mob) and desktop web domains (Desk). 95% CI designates the binomial proportion confidence interval for the accuracy at the 95% level assuming a normal distribution. The F-1 score based on features from both types of data (Mob & Desk - 108 Features) is higher than the scores obtained using mobile and desktop data individually (even with more features as in Desk - 2,923 Features). We observed similar results for value shoppers with F-1 scores of 0.17 (Mob - 85 Features), 0.25 (Desk - 99 Features), and 0.41 (Mob & Desk - 104 Features).

In particular, predicting from mobile data alone proved to be the weakest option. One reason seems to be that we only had 90 features from the mobile data compared to 106 and 107 for the desktop and combined data, respectively. Using desktop data only we tried to increase the performance to the level of the combined mobile and desktop data, which reached an F-1 score of 0.65. However, we were only able to obtain an F-1 score of 0.58 by substantially increasing the feature space to 2,923 features, at which point we saw no further improvement. Combining desktop and mobile data and leveraging 107 features outperformed all other approaches. The ROC curves in Figure 7 visualize this finding. The predictions that users have an interest in finance are shown in orange while the negative predictions for not having an interest in finance are displayed in blue. For the latter the F-1 scores are: Mob - 91 Features: 0.77, Desk - 107 Features: 0.83, Desk - 2,923 Features: 0.86, and Mob & Desk - 108 Features: 0.89.

Predicting Gender. While the predictive performance of a user's interest in finance increased with the availability of both desktop and mobile data, it appears that such improvement does not necessarily hold for all classification tasks. Particularly, classifier performance for the prediction of gender from combined desktop and mobile data was not better than the performance using desktop data alone. Applying logistic regression with 10-fold cross validation we obtained identical scores for precision, recall, and F-1 with values of 0.82, 0.81, and 0.82, respectively. It did not make a difference whether mobile data was added to the desktop data or not. This result suggests that for some tasks the availability of cross-device data does not lead to better predictions.

Impact of Device Usage Patterns. What could be the reason for the differing utility of cross-device data in the two prediction tasks? Subject to the results of further experiments

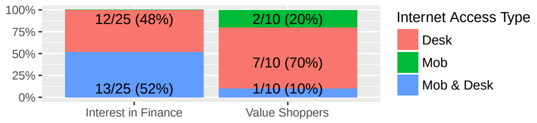


Figure 8: Mobile and desktop device usage patterns. Some users in our dataset access finance and value shopping domains only from their desktop or mobile device (i.e., from a mobile website or app).

it seems that having both mobile and desktop data available can be an advantage for predictions that rely on features exhibited on one device type only. We did not only observe such patterns for users with an interest in finance but also for value shoppers—a randomly selected persona category. For both interest in finance and the value shopper persona we evaluated to which extent users respectively accessed the top 25 finance and value shopping domains on their mobile and desktop devices. Our results, illustrated in Figure 8, support the conclusion that having data available from both mobile and desktop devices increases the chances of capturing (more) salient features for the aforementioned predictions. For example, an ad network without access to desktop data would have difficulty to make correct classifications for users that only access respective domains on their desktop device. We note that the observed patterns are based on a small number of users. Thus, further investigation is warranted.

Absence of Device Usage During our study we realized the possibility of making predictions about users who do not make use of their devices. Predicting a user's Jewish religion serves as an illustrative example.[12] Obviously, religious web domains and apps can be meaningful features for predicting adherence to a particular faith. However, such predictions are also possible based on subtler user behaviors. Most notably, as the data collection of our study covered the last two days of the Jewish Passover holiday we noticed that a few users in our study did not use either of their devices as the Jewish faith prescribes abstinence from using electronics. Among all users in our study the pattern of holiday observation became obvious. This signal was especially clear from the insight into multiple devices. While some users did not use one of their devices, only those observant of Passover did not use both. This example illustrates that device activity as such can be a useful predictor that might be exploitable by cross-device tracking.

7 The Scope of Cross-device Tracking

The scope of cross-device tracking on the Internet is yet to be explored. For example, through their integration into many websites and apps Facebook and Google appear to have vast reach into the various devices of their users as well as the ability to deterministically match those [73]. However, the percentage to which a typical Internet user is tracked across devices by those and other cross-device companies is not known. We examine this question for the users in our dataset based on a procedure for detecting the presence of cross-device trackers in their browsing and app histories (§ 7.1) and analyzing their occurrence accounting for industry collaborations and consolidation (§ 7.2).

7.1 Detecting Cross-device Trackers

Procedure. We examined the trackers on the websites that the users in our study visited by automating a Firefox browser with Selenium [74]. The browser included Lightbeam [63] and User Agent Switcher [64] browser extensions that allowed us to record the trackers on each domain for both mobile and desktop websites. Third party trackers that we found in a subdomain were added to the domain, however, not vice versa. Thus, for example, the domain linkedin.com contains all trackers that we found on blog.linkedin.com but not the other way around. To identify trackers inside of apps we selected a total of 153 third party software development kits (SDKs) listed on AppBrain [6] encompassing SDKs of ad networks (e.g., Smaato), social networks (e.g., Twitter), and analytics services (e.g., comScore). Leveraging AppBrain's statistics on the inclusion of SDKs in apps we then determined which SDKs are included in the apps of our dataset.

We qualify a company as cross-device company based on our detection of their trackers on both mobile and desktop domains, the former including apps, and their websites' claims that they indeed perform cross-device tracking. We identified cross-device trackers by using Whois domain searches and tracker blocking lists, especially, the list of the Better tracker blocker [42]. For some companies—Google AdSense, Rubicon Project, Skimlinks, Tapad, and Lotame—our information flow experiment (§ 3) provides empirical support for qualifying them as cross-device companies. It would be interesting to extend this or a similar information flow experiment towards other companies.

Lower Bound. Our approach for detecting trackers should be understood as a lower bound for various reasons. First, trackers not identified in Lightbeam will remain undetected. The same is true for SDKs not included in the pre-defined set of 153 SDKs from AppBrain. Further, we only detect app tracking via SDKs and do not account for Android WebViews and app-internal browsers that could contain tracking cookies or other traditional online tracking mechanisms [18]. We believe that these technologies warrant substantial further investigation as we suspect that a large amount of trackers make use of them.

Limitations. It is a limitation of our crawl that some websites in our dataset were not accessible (e.g., sites that required a user login). In some cases our crawl was also redirected or the requested page was not found. However, these limitations only affected few URLs. Also, it should

[12] As we obtained this result by chance we confirmed that its publication is covered by applicable IRB regulations of Columbia University.

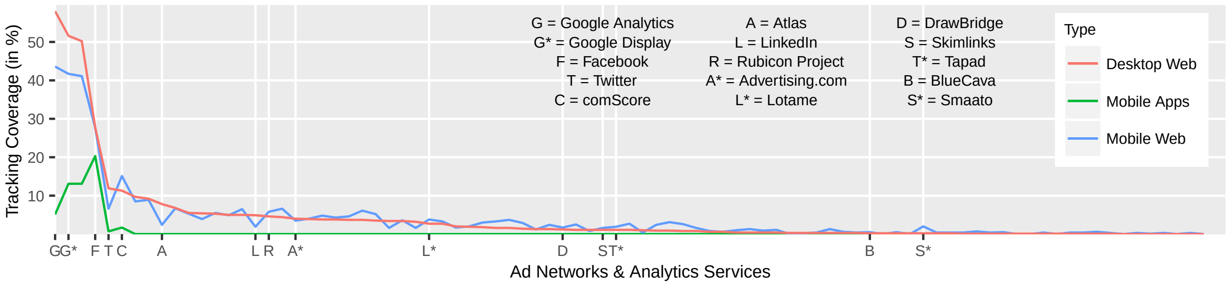


Figure 9: As the ad ecosystem in general, cross-device tracking is characterized by a few large companies with extensive reach and a long tail of smaller companies, some of which are focusing on the mobile space (e.g., Smaato with 2% mobile and 0.2% desktop coverage).

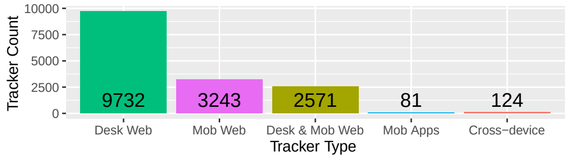


Figure 10: Total unique third party trackers in our dataset. We found 124 cross-device trackers that belong to 87 different companies.

be noted that we crawled the sites about a month after we finished collecting data from the study participants. Thus, in the meantime, some websites might have different trackers than at the time they were actually visited. Ideally, it would have been possible to capture the trackers live from our users' devices during the study. However, such collection is difficult due to the constraints of the Android environment, most notably, the sandboxing of mobile browsers. In addition, our mobile tracker count may be off as we did not use real mobile devices but instead a spoofed desktop browser.

7.2 Cross-Device Tracking Analysis

As shown in Figure 10, websites accessed from desktop and mobile devices contained a respective total of 9,732 and 3,243 unique third party trackers. 2,571 trackers were on both desktop and mobile websites; Brookman et al. [11] found 861 such trackers. Out of the 153 SDKs from AppBrain we found 81 in our dataset.[13] From these sets of third party trackers we identified 124 cross-device trackers; 118 trackers that appeared on both mobile and desktop websites and 6 SDKs that are associated with a desktop tracker as well. We found that the 124 cross-device trackers belong to 87 different companies. It appears that 22 follow a deterministic approach, 39 use probabilistic techniques, and 26 leverage both.

Tracking of the Average User in our Dataset. On average each user is tracked across devices on his or her desktop in 67% of all desktop website visits. We measured a similar average for mobile web visits with 64%. These high percentages illustrate that cross-device tracking is a broadly occurring phenomenon. Table 5 shows the reach of individual companies. Google Analytics, Google Display, and Facebook can capture at least 20% of an average user's online traffic across devices. This percentage is about the same that Roesner et al. [72] provided a few years ago for tracking of individual devices. It is particularly noteworthy that the companies with the broadest reach have a deterministic approach, which means that their cross-device tracking is also very accurate. Figure 9 shows the tracking coverage for the 87 cross-device companies we identified.

	Desk	*Mob*	*Apps*
Google Analytics (D)	58%	43.6%	5.1%
Google Display (D,P)	51.6%	41.7%	13.1%
Facebook (D)	27.8%	27.7%	20.3%
Atlas (Facebook) (D,P)	7.8%	2.4%	N/A*
Facebook & Atlas (D,P)	**27.9%**	**29.1%**	**20.3%**
Twitter (D)	11.9%	6.6%	0.7%
comScore (D,P)	11.3%	15.1%	1.7%
LinkedIn (Microsoft) (D)	4.9%	1.9%	N/A*
Rubicon Project (P)	4.6%	5.8%	N/A*
Tapad (P)	1.1%	1.9%	N/A
Rubicon & Tapad (P)	5.4%	6.7%	N/A*
Advertising.com (AOL) (P)	4%	3.5%	N/A
Lotame (D,P)	2.7%	3.8%	N/A*
Skimlinks (D,P)	1.1%	1.6%	N/A
Lotame & Skimlinks (D,P)	**3.5%**	**5%**	N/A*
Drawbridge (P)	1.2%	1.7%	N/A
BlueCava (P)	0.2%	0.5%	N/A
Smaato (D,P)	**0.2%**	**2%**	**0.1%**

Table 5: Cross-device companies' (D = deterministic and P = probabilistic according to their websites' claims) coverage of websites and apps on average for the users in our dataset (n = 107). Some of the companies either do not seem to offer an SDK for app integration (N/A) or we did not analyze it as it was not contained in our initial set of SDKs from AppBrain (N/A*). The full list, including the tracking server domains, is attached in Appendix B.

Partnerships between various cross-device companies extend their reach. For example, Atlas receives user data from Facebook [8] to track users deterministically. However, Atlas

[13] The app tracker count includes affiliated company's SDKs. Thus, for example, the Facebook SDK inside the Instagram app is counted as one tracker.

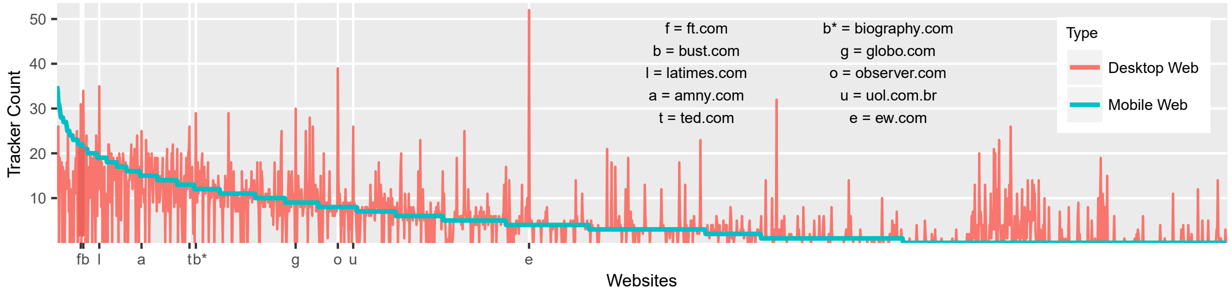


Figure 11: The ten domains with the highest number of cross-device companies on their desktop websites (out of 1,829 total domains). It can be observed that they tend to have higher concentrations of cross-device companies on their mobile sites as well.

	Desk	*Mob*	*Rank-Country*
ew.com	**52**	4	466-US
observer.com	**39**	8	1,191-US
latimes.com	**35**	19	133-US
bust.com	**34**	21	25,690-US
ft.com	**31**	22	166-UK
globo.com	**30**	9	5-BR
biography.com	**29**	12	1141-US
ted.com	**26**	13	635-US
uol.com.br	**26**	8	N/A
amny.com	**25**	15	N/A
gameofthrones.wikia.com	14	**35**	45-US
androidauthority.com	11	**34**	570-IN
food.com	**26**	**31**	600-US
sacbee.com	0	**31**	2,985-US
sfgate.com	19	**30**	310-US
philly.com	15	**29**	908-US
southcoasttoday.com	17	**28**	N/A
nypost.com	17	**28**	154-US
nytimes.com	18	**28**	48-US
jalopnik.com	15	**27**	782-US

Table 6: Domains with the highest cross-device company counts out of 1,829 domains whose URL occurred in both mobile and desktop data in our CDT dataset. With a total of 57 trackers (31 mobile web and 26 desktop web) food.com had the highest count overall.

is intended to serve advertisements outside of Facebook's reach and shares the data it collects with Facebook as well [8]. Particularly, as shown in Table 5, Atlas' cross-device trackers extend Facebook's mobile web reach from 27.7% to 29.1%. As another example, the partnership between Lotame and Skimlinks [57], which we actually observed in our initial experiment (§ 3), also extends their respective reach. In those cases the relationship between companies needs to be accounted for to accurately determine their full coverage.

Domains with Cross-device Company Concentration. It appears that media websites, in particular, websites of newspapers, contain the largest concentration of trackers from cross-device companies. Table 6 shows the top ten domains—separated for desktop and mobile websites—on which we found the highest number of trackers from the 87 identified cross-device companies. Coincidentally, it turns out that the website of the LA Times was a good selection for our case study (§ 3) as it had trackers from 35 cross-device companies on its desktop website.

Beyond the concentration of cross-device companies' trackers in the media category it is also striking that many websites that are hosting those trackers are fairly popular sites. Table 6 shows the Quantcast country rank according to the site's traffic [70]. This placement of cross-device trackers on popular sites exposes them to large audiences. However, as it can be observed in Table 6 as well, the shown domains contain a maximum number of trackers from cross-device companies on either their mobile or desktop sites but not on both. This finding holds in general. While there is a tendency that domains that host many cross-device companies on their desktop site also host many on their mobile site, we could not find any statistically significant correlation in this regard. Figure 11 shows the distribution.

8 Does Self-Regulation Work?

The FTC recommends that cross-device companies should be transparent about their data practices [32]. While there are no specific statutes or regulations for cross-device tracking in the U.S., the field is subject to self-regulation, most notably by the Digital Advertising Alliance (DAA) and the Network Advertising Initiative. The DAA requires its member cross-device companies to disclose "the fact that data collected from a particular browser or device may be used with another computer or device that is linked to the browser or device on which such data was collected." [21] In order to examine the level of compliance with this transparency requirement we randomly selected 40 DAA member ad networks that advertised their cross-device capabilities on their websites and analyzed their privacy policies.

We found that 23 disclosed their cross-device tracking activities while 17 omitted those. After contacting the latter, we received a response from seven. Two pointed us to documents that were linked from the policy that indeed contained compliant descriptions. A representative from another cross-device company wrote that their cross-device functionality is not yet fully rolled out to clients, and three others announced that they will change their policy (one of those still has to follow through). Another representative simply claimed that the company is “not violating anything.” Without contacting us five further cross-device companies simply changed their policies, of which four became compliant. Finally, there was no reaction or policy change from five. As of June 9, 2017 we count a total of eight instances of non-compliance.

Overall, it appears that there is a lack of transparency when it comes to the disclosure of cross-device tracking. At this point, the DAA guidance does not seem to be enforced rigorously. While it may be true that the majority of consumers will not take the time to understand the tracking practices described in privacy policies,[14] we think that it is still a worthwhile endeavor for cross-device companies to properly disclose their practices, particularly, for audit and enforcement purposes as well as for signaling trustworthiness to the marketplace and to build an environment of rules and norms in privacy disclosure.

9 Conclusion

Cross-device tracking is an emerging tracking paradigm that challenges current notions of privacy. This study is intended as a broad overview of selected privacy topics in mainstream cross-device technologies. In a brief case study we have demonstrated how cross-device tracking can be observed with statistical confidence by means of an information flow experiment. Using our own cross-device tracking dataset we designed a cross-device tracking algorithm and evaluated relevant features and parameter settings grounded in a review of publicly available information on the practices of cross-device companies. For some predictive tasks it appears that those companies can learn more about users than from individual device data. As the penetration of cross-device tracking on the Internet already appears relatively high it is even more important that companies active in this field are transparent about their practices.

Going forward we hope that the various privacy implications of cross-device tracking technologies will be studied further. In this regard, proprietary research is substantially ahead of current efforts in academia. While a few major points are known—for example, that IP addresses are a crucial feature for correlating devices—many important details on how cross-device companies operate remain opaque. To shed more light on the subject we publicized our dataset together with the software that we developed for further exploration. As cross-device tracking continues to mature and become an integral part of tracking on the Internet we believe that a comprehensive view including legal and business considerations is helpful. Establishing an enforceable self-regulatory framework for companies to be transparent about their practices will help to protect consumer privacy and allow cross-device companies to conduct their businesses responsibly.

Ultimately, cross-device tracking is part of a larger trend: the Internet of Things (IoT). In this regard, we see cross-device tracking as an early harbinger of the increasing inter-connectivity of devices. Increasingly, buildings, cars, appliances, and other things are connected to the Internet and are interacting with other online devices. However, the development and deployment of privacy solutions has to keep pace with the emerging IoT landscape. Ensuring transparency and practicable control mechanisms for information that is traversing device boundaries and permeates between the online and offline worlds is a critical element. Given standardized interfaces [43], perhaps, an intelligent personal privacy assistant that is connected to all services and devices of person could be a solution.

Acknowledgment

We would like to thank the anonymous reviewers as well as our shepherd Alina Oprea for their helpful comments during the paper review phase. We further thank Clément Canonne, Paul Blaer, Daisy Nguyen, and Anupam Das. Tony Jebara is supported in part by DARPA (N66001-15-C-4032) and NSF (III-1526914, IIS-1451500). We also gratefully acknowledge the Comcast Research Grant for our user study. The views and conclusions contained herein are our own.

References

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

[2] Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F., and Preneel, B. FPDetective: Dusting the web for fingerprinters. In *CCS 2013*, ACM.

[3] Adar, E., Teevan, J., and Dumais, S. T. Large scale analysis of web revisitation patterns. In *CHI 2008*, ACM.

[4] Alexa. The top 500 sites on the web. `http://www.alexa.com/topsites/countries/US`. Accessed: June 29, 2017.

[5] Alexa. The top 500 sites on the web. `http://www.alexa.com/topsites/category`. Accessed: June 29, 2017.

[6] AppBrain. Android library statistics. `http://www.appbrain.com/stats/libraries/`. Accessed: June 29, 2017.

[7] Arp, D., Quiring, E., Wressnegger, C., and Rieck, K. Privacy threats through ultrasonic side channels on mobile devices. In *EuroS&P 2017*, IEEE Computer Society.

[8] Atlas. Privacy policy. `https://atlassolutions.com/privacy-policy/`, Apr. 2015. Accessed: June 29, 2017.

[9] BlueCava, Inc. `http://bluecava.com/`. Accessed: June 29, 2017.

[14] Using tracking protection software and ad blockers is a much more efficient approach from a user perspective. Thus, when evaluating cross-device tracking in terms of a threat model, the most effective defense would be to block tracking. In this regard, the defenses against cross-device tracking are the same as the defenses against the tracking of individual devices.

[10] Book, T., and Wallach, D. S. An empirical study of mobile ad targeting. *CoRR abs/1502.06577* (2015).

[11] Brookman, J., Rouge, P., Alva, A., and Yeung, C. Cross-device tracking: Measurement and disclosures. In *PoPETs 2017*, De Gruyter.

[12] Cai, X., Nithyanand, R., and Johnson, R. CS-BuFLO: A congestion sensitive website fingerprinting defense. In *WPES 2014*, ACM.

[13] Cai, X., Nithyanand, R., Wang, T., Johnson, R., and Goldberg, I. A systematic approach to developing and evaluating website fingerprinting defenses. In *CCS 2014*, ACM.

[14] Cao, X., Huang, W., and Yu, Y. Recovering cross-device connections via mining IP footprints with ensemble learning. In *IEEE International Conference on Data Mining Workshop, ICDMW 2015*.

[15] Cao, Y., Li, S., and Wijmansy, E. (Cross-)browser fingerprinting via os and hardware level features. In *NDSS 2016*, Internet Society.

[16] Chao, A., and Shen, T.-J. Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics 10*, 4 (2003), 429–443.

[17] Cherubin, G., Hayes, J., and Juarez, M. Website fingerprinting defenses at the application layer. In *PoPETs 2017*, De Gruyter.

[18] Criteo SA. Building the cross device graph at criteo. `http://labs.criteo.com/2016/06/building-cross-device-graph-criteo/`. Accessed: June 29, 2017.

[19] Das, A., Borisov, N., and Caesar, M. Tracking mobile web users through motion sensors: Attacks and defenses. In *NDSS 2016*, Internet Society.

[20] Dearman, D., and Pierce, J. S. It's on my other computer!: Computing with multiple devices. In *CHI 2008*, ACM.

[21] Digital Advertising Alliance. Application of the self-regulatory principles of transparency and control to data used across devices. `http://www.aboutads.info/sites/default/files/DAA_Cross-Device_Guidance-Final.pdf`, Nov. 2015. Accessed: June 29, 2017.

[22] Drawbridge, Inc. `http://www.drawbrid.ge/`. Accessed: June 29, 2017.

[23] Drawbridge, Inc. Drawbridge challenges scientific community to better the accuracy of its cross-device consumer graph. `https://drawbridge.com/news/p/drawbridge-challenges-scientific-community-to-better-the-accuracy-of-its-cross-device-consumer-graph`. Accessed: June 29, 2017.

[24] Dstillery, Inc. A tale of two crosswalks. `http://dstillery.com/a-tale-of-two-crosswalks/`. Accessed: June 29, 2017.

[25] Eckersley, P. How unique is your web browser? In *PETS 2010*, Springer-Verlag.

[26] Englehardt, S., and Narayanan, A. Online tracking: A 1-million-site measurement and analysis. In *CCS 2016*, ACM.

[27] Englehardt, S., Reisman, D., Eubank, C., Zimmerman, P., Mayer, J., Narayanan, A., and Felten, E. W. Cookies that give you away: The surveillance implications of web tracking. In *WWW 2015*, International World Wide Web Conferences Steering Committee.

[28] Experian Ltd. Device recognition by adtruth. `http://www.experian.co.uk/marketing-services/products/adtruth-device-recognition.html`. Accessed: June 29, 2017.

[29] Federal Trade Commission. FTC cross-device tracking workshop. `https://www.ftc.gov/news-events/events-calendar/2015/11/cross-device-tracking`, Nov. 2015. Accessed: June 29, 2017.

[30] Federal Trade Commission. FTC cross-device tracking workshop, segment 1, transcript. `https://www.ftc.gov/system/files/documents/videos/cross-device-tracking-part-1/ftc_cross_device_tracking_workshop_-_transcript_segment_1.pdf`, Nov. 2015. Accessed: June 29, 2017.

[31] Federal Trade Commission. FTC issues warning letters to app developers using 'Silverpush' code. `https://www.ftc.gov/news-events/press-releases/2016/03/ftc-issues-warning-letters-app-developers-using-silverpush-code`, Mar. 2016. Accessed: June 29, 2017.

[32] Federal Trade Commission. Cross-device tracking. An FTC staff report. `https://www.ftc.gov/reports/cross-device-tracking-federal-trade-commission-staff-report-january-2017`, Jan. 2017. Accessed: June 29, 2017.

[33] Gesellschaft für Konsumforschung. Finding simplicity in a multi-device world. `https://blog.gfk.com/2014/03/finding-simplicity-in-a-multi-device-world/`, Mar. 2014. Accessed: June 29, 2017.

[34] Google Display Network. Where ads might appear in the display network. `https://support.google.com/adwords/answer/2404191?hl=en`. Accessed: June 29, 2017.

[35] Google, Inc. General ad categories. `https://support.google.com/adsense/answer/3016459?hl=en`. Accessed: June 29, 2017.

[36] Google, Inc. Topics used for personalized ads. `https://support.google.com/ads/answer/2842480?hl=en`. Accessed: June 29, 2017.

[37] Google, Inc. The new multi-screen world study. `https://www.thinkwithgoogle.com/research-studies/the-new-multi-screen-world-study.html`, Aug. 2012. Accessed: June 29, 2017.

[38] Google Play store. `https://play.google.com/store/apps?hl=en`. Accessed: June 29, 2017.

[39] Gulyás, G. G., Ács, G., and Castelluccia, C. Near-optimal fingerprinting with constraints. In *PoPETs 2016*, De Gruyter.

[40] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA data mining software: An update. *SIGKDD Explor. Newsl. 11*, 1 (Nov. 2009), 10–18.

[41] Hayes, J., and Danezis, G. k-fingerprinting: A robust scalable website fingerprinting technique. In *USENIX Security 2016*, USENIX Association.

[42] Ind.ie. Better tracker blocker. `https://better.fyi/trackers/`. Accessed: June 29, 2017.

[43] INFSO D.4 Networked Enterprise & RFID IMFSO G.2 Micro & Nanosystems, RFID Working Group of the European Technology Platform on Smart Systems Integration (EPOSS). Internet of things in 2020. `http://www.smart-systems-integration.org/public/documents/publications/Internet-of-Things_in_2020_EC-EPoSS_Workshop_Report_2008_v3.pdf`, Sept. 2008. Accessed: June 29, 2017.

[44] Juarez, M., Afroz, S., Acar, G., Diaz, C., and Greenstadt, R. A critical evaluation of website fingerprinting attacks. In *CCS 2014*, ACM.

[45] Kaggle, Inc. ICDM 2015: Drawbridge cross-device connections. `https://www.kaggle.com/c/icdm-2015-drawbridge-cross-device-connections/data`. Accessed: June 29, 2017.

[46] Kamvar, M., Kellar, M., Patel, R., and Xu, Y. Computers and Iphones and mobile phones, oh my!: A logs-based comparison of search users on different devices. In *WWW 2009*, International World Wide Web Conferences Steering Committee.

[47] KANE, S. K., KARLSON, A. K., MEYERS, B. R., JOHNS, P., JACOBS, A., AND SMITH, G. *Exploring Cross-Device Web Use on PCs and Mobile Devices*. Springer-Verlag, 2009, pp. 722–735.

[48] KEJELA, G., AND RONG, C. Cross-device consumer identification. In *IEEE International Conference on Data Mining Workshop, ICDMW 2015*.

[49] KIHN, M. Cross-device identity: A data scientist speaks. `http://blogs.gartner.com/martin-kihn/cross-device-identity-a-data-scientist-speaks/`, Oct. 2016. Accessed: June 29, 2017.

[50] KIM, M. S., LIU, J., WANG, X., AND YANG, W. Connecting devices to cookies via filtering, feature engineering, and boosting. In *IEEE International Conference on Data Mining Workshop, ICDMW 2015*.

[51] KOHNO, T., BROIDO, A., AND CLAFFY, K. Remote physical device fingerprinting. *IEEE Transactions on Dependable and Secure Computing 2*, 2.

[52] KURTZ, A., GASCON, H., BECKER, T., RIECK, K., AND FREILING, F. C. Fingerprinting mobile devices using personalized configurations. In *PoPETs 2016*, De Gruyter.

[53] LANDRY, M., S, S. R., AND CHONG, R. Multi-layer classification: ICDM 2015 drawbridge cross-device connections competition. In *IEEE International Conference on Data Mining Workshop, ICDMW 2015*.

[54] LÉCUYER, M., DUCOFFE, G., LAN, F., PAPANCEA, A., PETSIOS, T., SPAHN, R., CHAINTREAU, A., AND GEAMBASU, R. Xray: Enhancing the web's transparency with differential correlation. In *USENIX Security 2014*, USENIX Association.

[55] LÉCUYER, M., SPAHN, R., SPILIOPOLOUS, Y., CHAINTREAU, A., GEAMBASU, R., AND HSU, D. J. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *CCS 2015*, ACM.

[56] LERNER, A., SIMPSON, A. K., KOHNO, T., AND ROESNER, F. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *USENIX Security 2016*, USENIX Association.

[57] LOTAME SOLUTIONS, INC. Skimlinks and Lotame unleash enhanced retail intent data. `https://www.lotame.com/resource/skimlinks-lotame-dmp/`. Accessed: June 29, 2017.

[58] MATTHEWS, T., LIAO, K., TURNER, A., BERKOVICH, M., REEDER, R., AND CONSOLVO, S. "She'll just grab any device that's closer": A study of everyday device & account sharing in households. In *Proceedings of the ACM Conference on Human Factors in Computing Systems 2016*.

[59] MAVROUDIS, V., HAO, S., FRATANTONIO, Y., MAGGI, F., VIGNA, G., AND KRUEGEL, C. On the privacy and security of the ultrasound ecosystem. In *PoPETs 2017*, De Gruyter.

[60] MENG, W., DING, R., CHUNG, S. P., HAN, S., AND LEE, W. The price of free: Privacy leakage in personalized mobile in-apps ads. In *NDSS 2016*, Internet Society.

[61] MENG, W., XING, X., SHETH, A., WEINSBERG, U., AND LEE, W. Your online interests: Pwned! a pollution attack against targeted advertising. In *CCS 2014*, ACM.

[62] MORALES, R. D. Cross-device tracking: Matching devices and cookies. In *IEEE International Conference on Data Mining Workshop, ICDMW 2015*.

[63] MOZILLA. Lightbeam for Firefox. `https://www.mozilla.org/en-US/lightbeam/`. Accessed: June 29, 2017.

[64] MYBROWSERADDON. User-agent switcher. `http://mybrowseraddon.com/useragent-switcher.html`. Accessed: June 29, 2017.

[65] NIKIFORAKIS, N., JOOSEN, W., AND LIVSHITS, B. Privaricator: Deceiving fingerprinters with little white lies. In *WWW 2015*, International World Wide Web Conferences Steering Committee.

[66] NIKIFORAKIS, N., KAPRAVELOS, A., JOOSEN, W., KRUEGEL, C., PIESSENS, F., AND VIGNA, G. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *S&P 2013*, IEEE Computer Society.

[67] OLEJNIK, L., CASTELLUCCIA, C., AND JANC, A. On the uniqueness of web browsing history patterns. *Annales des Télécommunications 69*, 1-2 (2014), 63–74.

[68] PAGEFAIR. Adblocking goes mobile. `https://pagefair.com/downloads/2016/05/Adblocking-Goes-Mobile.pdf`. Accessed: June 29, 2017.

[69] PANCHENKO, A., LANZE, F., ZINNEN, A., HENZE, M., PENNEKAMP, J., WEHRLE, K., AND ENGEL, T. Website fingerprinting at internet scale. In *NDSS 2016*, Internet Society.

[70] QUANTCAST. Top sites. `https://www.quantcast.com/top-sites`. Accessed: June 29, 2017.

[71] RENOV, O., AND ANAND, T. R. Machine learning approach to identify users across their digital devices. In *IEEE International Conference on Data Mining Workshop, ICDMW 2015*.

[72] ROESNER, F., KOHNO, T., AND WETHERALL, D. Detecting and defending against third-party tracking on the web. In *NSDI 2012*, USENIX Association.

[73] SCHIFF, A. 2016 edition: A marketers guide to cross-device identity. `https://adexchanger.com/data-exchanges/2016-edition-marketers-guide-cross-device-identity/`, Feb. 2016. Accessed: June 29, 2017.

[74] SELENIUMHQ. Seleniumhq browser automation. `http://www.seleniumhq.org/`. Accessed: June 29, 2017.

[75] SELSAAS, L. R., AGRAWAL, B., RONG, C., AND WIKTORSKI, T. AFFM: auto feature engineering in field-aware factorization machines for predictive analytics. In *IEEE International Conference on Data Mining Workshop, ICDMW 2015*.

[76] TAPAD, INC. `http://www.tapad.com/`. Accessed: June 29, 2017.

[77] TAPAD, INC. Nielsen confirms tapad cross-device accuracy at 91.2%. `http://www.tapad.com/news/blog/nielsen-confirms-tapad-cross-device-accuracy-at-91-2`, Dec. 2014. Accessed: June 29, 2017.

[78] TAUSCHER, L., AND GREENBERG, S. How people revisit web pages. *Int. J. Hum.-Comput. Stud. 47*, 1 (July 1997), 97–137.

[79] TOSSELL, C., KORTUM, P., RAHMATI, A., SHEPARD, C., AND ZHONG, L. Characterizing web use on smartphones. In *CHI 2012*, ACM.

[80] TRAASDAHL, A., LIODDEN, D., AND CHANG, V. Managing associations between device identifiers, May 16 2013. US Patent App. 13/677,110.

[81] TSCHANTZ, M. C., DATTA, A., DATTA, A., AND WING, J. M. A methodology for information flow experiments. In *CSF 2015*, IEEE Computer Society.

[82] WALTHERS, J. Learning to rank for cross-device identification. In *IEEE International Conference on Data Mining Workshop, ICDMW 2015*.

[83] WANG, C., AND PU, H. Uniquely identifying a network-connected entity, May 7 2013. US Patent 8,438,184.

[84] YAHOO! INC. Personas. `https://web.archive.org/web/20160728012832/https://developer.yahoo.com/flurry/docs/analytics/lexicon/personas/`. Accessed: June 29, 2017.

[85] ZARRAS, A., KAPRAVELOS, A., STRINGHINI, G., HOLZ, T., KRUEGEL, C., AND VIGNA, G. The dark alleys of madison avenue: Understanding malicious advertisements. In *IMC 2014*, ACM.

A Cross-device Tracking Dataset

***Device Fingerprints** (n = 234)*

User Agent	1st Party HTTP Cookies Enabled
Browser Vendor	3rd Party HTTP Cookies Enabled
Browser Engine	Do Not Track Enabled
Plugins Installed	Touchscreen
Operating System	Internet Connection Type
Time Zone	Latency
Screen (Color Depth, etc.)	Fonts Installed
System Language	Local Storage Enabled
Adobe Flash Version	Session Storage Enabled
Microsoft Silverlight Version	HTTP Accept Headers
JavaScript Enabled	

***App and Browsing Histories** (n = 233)*

IP Address	Browser Tab ID
Browser Vendor	Referrer URL
Date	URL/App Package ID
Time	URL Title
Time Zone	3rd Party Trackers/SDKs

***Interest Questionnaires** (n = 126)*

Arts and Entertainment (68%)	Beauty and Fitness (33%)
Food and Drink (64%)	Internet and Telecom (33%)
Computers and Electronics (63%)	Sports (29%)
Science (62%)	Online Communities (24%)
News (60%)	Finance (23%)
Books and Literature (55%)	Pets and Animals (23%)
Jobs and Education (52%)	Business and Industrial (21%)
Games (43%)	World Localities (15%)
Travel (40%)	Reference (13%)
Law and Government (37%)	Autos and Vehicles (11%)
Shopping (36%)	Home and Garden (11%)
Hobbies and Leisure (34%)	Real Estate (4%)
People and Society (34%)	

***Persona Questionnaires** (n = 126)*

Music Lovers (47%)	Hardcore Gamers (11%)
Movie Lovers (46%)	Photo and Video Enthusiasts (11%)
Food and Dining Lovers (40%)	Fashionistas (10%)
Singles (39%)	Personal Finance Geeks (10%)
Bookworms (33%)	Avid Runners (7%)
Entertainment Enthusiasts (31%)	Flight Intenders (6%)
Tech and Gadget Enthusiasts (31%)	Social Influencers (6%)
Casual and Social Gamers (30%)	Catalog Shoppers (5%)
News and Magazine Readers (23%)	Auto Enthusiasts (3%)
Leisure Travelers (21%)	Business Travelers (3%)
Sports Fans (21%)	Small Business Owners (3%)
Health and Fitness Enthusiasts (20%)	Home Design Enthusiasts (2%)
Mobile Payment Makers (19%)	Real Estate Followers (2%)
Value Shoppers (18%)	High Net Individuals (1%)
Parenting and Education (15%)	Mothers (1%)
Pet Owners (14%)	Home and Garden Pros (0%)
Business Professionals (13%)	New Mothers (0%)
American Football Fans (11%)	Slots Players (0%)

***Age Groups** (n = 126)*

18–20 (18%)	31–35 (6%)
21–25 (51%)	over 35 (3%)
26–30 (21%)	

***Gender** (n = 126)*

Women (34%)	Men (66%)

B Cross-device Trackers

	Type	*Desk Web*	*Mob Web*	*Mob Apps*
33Across - 33across.com	P	0.1	0.4	N/A
Adbrain - adbrn.com	P	0.4	1	N/A
AddThis (Oracle) - addthis.com - addthisedge.com	P	3.4	1.6	N/A*
Adelphic - ipredictive.com	D, P	0.1	0.7	N/A
Adform - adform.net - adformdsp.net	D, P	0.4	1.3	N/A*
Adobe Marketing Cloud - 2o7.net - adobetag.com - omtrdc.net	D	5.4	3.9	N/A*
AdRoll - adroll.com	D	2	1.7	N/A
Advertising.com (AOL) - advertising.com	P	4	3.5	N/A
Amobee - amgdgt.com	D	0	0	N/A*
AOL ONE - adap.tv - adtech.de - adtechus.com - aol.com - atwola.com - jumptap.com	P	3.7	4.6	N/A*
AppNexus - adnxs.com	P	9.2	8.9	N/A*
Arbor - pippio.com	D	0.1	0.4	N/A
Atlas (Facebook) - atdmt.com	D, P	7.8	2.4	N/A*
AudienceScience - revsci.net	D, P	0.9	2.4	N/A
Baidu - baidu.com	P	0.3	0	N/A
Bidtellect - bttrack.com	P	0.1	0.5	N/A
Bing ads (Microsoft) - bing.com - msn.com	D	3.2	1.6	N/A
BlueCava - bluecava.com	P	0.2	0.5	N/A
BlueKai (Oracle) - bkrtx.com - bluekai.com	D, P	3.8	4.8	N/A*
BrightRoll (Yahoo) - btrll.com	D, P	0.8	2.6	N/A
Cardlytics - cardlytics.com	P	0.2	0	N/A
Casale Media - indexww.com - casalemedia.com	P	3.8	4.3	N/A
ChoiceStream - choicestream.com	D, P	0.1	0	N/A
Clearstream TV - clrstm.com	P	0.1	0	N/A
comScore - comscore.com - scorecardresearch.com - zqtk.net - comScore SDK	D, P	11.3	15.1	1.7
Connexity - connexity.net	P	0.1	0.4	N/A
Crimtan - ctnsnet.com	D, P	0	0.3	N/A
Criteo	D	5.3	5.5	N/A

	Type	*Desk Web*	*Mob Web*	*Mob Apps*
- criteo.com - criteo.net				
Cross Pixel Media - crsspxl.com	D	0.3	0.2	N/A
Datalogix (Oracle) - nexac.com	P	1.6	3.3	N/A
DataXu - w55c.net	D, P	1.1	2.5	N/A
Datonics - pro-market.net	P	0.2	0.5	N/A
Deep Forest Media (Rakuten) - dpclk.com	P	0.3	0.4	N/A
Demandbase - company-target.com	D	0.2	0	N/A
DistroScale - jsrdn.com	P	0.1	0	N/A
DoubleClick (Google) - 2mdn.net - dmtry.com - doubleclick.net - AdMob SDK	D, P	50.2	41.1	13.1
Drawbridge - adsymptotic.com	P	1.2	1.7	N/A
Dstillery - media6degrees.com	P	0.3	1.3	N/A
engage:BDR - bnmla.com	P	0	0.1	N/A
Ensighten - ensighten.com	D	1.3	1.2	N/A
eXelate (Nielsen) - exelator.com	D, P	1.4	2.9	N/A*
Eyereturn marketing - eyereturn.com	D	0	0.3	N/A
Eyeview - eyeviewads.com	P	0.1	0.4	N/A
Facebook - facebook.com - facebook.net - fb.me - Facebook SDK	D	27.8	27.7	20.3
FreeWheel (Comcast) - fwmrm.net	P	0.3	0.6	N/A
Gigya - gigya.com	D	0.6	0.8	N/A
Google Analytics - google-analytics.com - Google Analytics SDK	D	58	43.6	5.1
Google Display Network - 2mdn.net - adsense.com - blogger.com - dmtry.com - doubleclick.net - googleadservices.com - youtube.com - AdMob SDK	D, P	51.6	41.7	13.1
IgnitionOne - netmng.com	P	1	0.4	N/A
Interstate - interstateanalytics.com	D	0	0	N/A
IXI Services (Equifax) - ixiaa.com	D, P	1.6	3.7	N/A
Kenshoo - xg4ken.com	D, P	0.1	0.4	N/A
Krux - krxd.net	D, P	3.5	5.2	N/A
LinkedIn - bizographics.com - linkedin.com	D	4.9	1.9	N/A*
Lotame - crwdcntrl.net	D, P	2.7	3.8	N/A*
Magnetic - domdex.com	P	0.6	0.6	N/A
MaxPoint - mxptint.net	D	0.1	0.6	N/A
MediaMath - mathtag.com	P	5	4.9	N/A
Moat - moatads.com	D, P	6.8	6.7	N/A
Neustar - adadvisor.net - agkn.com	D, P	3.7	6.1	N/A
Nielsen - imrworldwide.com	D, P	5.5	5.3	N/A*
Optimizely - optimizely.com	D	3.4	3.6	N/A*
Perfect Audience - prfct.co	P, D	0.3	0.4	N/A*
PubMatic - pubmatic.com	P	2.7	3.3	N/A*
Quantcast - quantserve.com	P	9.7	8.5	N/A
RadiumOne - gwallet.com	D	0.4	0.9	N/A*
Resonate - reson8.com	P	0	0.3	N/A
Rocket Fuel - rfihub.com	P	1.8	3	N/A
Rubicon Project - chango.com - rubiconproject.com	P	4.6	5.8	N/A*
RUN - rundsp.com	P	0.1	0.3	N/A
Signal - thebrighttag.com	D	1.1	0.8	N/A
Sizmek - peer39.com - peer39.net - serving-sys.com	P	3.9	4	N/A
Skimlinks - skimresources.com - redirectingat.com	D, P	1.1	1.6	N/A
Smaato - smaato.net - Smaato SDK	D, P	0.2	2	0.1
Smart AdServer - smartadserver.com	P	0.4	1.1	N/A
Sonobi - sonobi.com	P	1.3	2.4	N/A
SpotX - spotxchange.com	D, P	0.9	3.1	N/A*
Tapad - tapad.com	P	1.1	1.9	N/A
Tealium - tiqcdn.com	D	1.9	2	N/A
The Trade Desk - adsrvr.org	P, D	4.4	6.6	N/A
Turn - turn.com	P	1.1	2.7	N/A
Twitter - ads-twitter.com - twitter.com - Twitter SDK	D	11.9	6.6	0.7
Undertone - legolas-media.com - undertone.com	P	0.2	0.4	N/A
Vindico (Time) - vindicosuite.com	D, P	0.2	0.4	N/A*
Weborama - weborama.fr - weborama.io	P	0	0	N/A*
Yahoo - yahoo.com	D	5	6.5	N/A*
Yieldbot - yldbt.com	P	0.8	1.6	N/A