Investigating the Privacy Implications of Cross-Device Tracking

Konstantinos Solomos

Thesis submitted in partial fulfillment of the requirements for the

Masters' of Science degree in Computer Science and Engineering

University of Crete
School of Sciences and Engineering
Computer Science Department
Voutes University Campus, 700 13 Heraklion, Crete, Greece

Thesis Advisors:
Prof. *Evangelos Markatos*
Dr. *Sotiris Ioannidis*

This work has been performed at the University of Crete, School of Sciences and Engineering, Computer Science Department.

The work has been supported by the Foundation for Research and Technology - Hellas (FORTH), Institute of Computer Science (ICS).

Investigating the Privacy Implications of Cross-Device Tracking

Abstract

Although digital advertising fuels much of today's free Web, it typically does so at the cost of online users' privacy, due to continuous tracking and leakage of users' personal data. In search for new ways to optimize the effectiveness of online ads, advertisers have introduced new advanced methods such as *Cross-Device Tracking* (CDT), to monitor users' browsing activity on multiple devices and screens, and deliver (re)targeted ads in the most appropriate screen. Unfortunately, this practice leads to greater privacy concerns for the end-user, not extensively studied before.

In this thesis, we propose a novel methodology for detecting and measuring Cross-Device Tracking, and investigating the factors affecting its performance in a repeatable and systematic way. This new methodology is based on emulating realistic browsing activity of end-users from different devices, and thus triggering, detecting and classifying cross-device targeted ads. We implement this methodology in a novel CDT measurement framework that allows experimentation with multiple parallel devices, setups and experimental configurations. By employing our framework, we perform several critical experiments, and we are able not only to detect and measure CDT with average accuracy of 78-96%, but also to provide significant insights about the behavior of CDT entities and the impact on users' privacy.

In fact, our modular and extensible design allows us to investigate Cross-Device Tracking in depth and propose new extensions to study the complex structure of the ad-ecosystem. Our findings can be useful for raising awareness and increasing transparency on tracking practices used by online advertisers.

Διερεύνηση των Επιπτώσεων στην Ιδιωτικότητα της Παρακολούθησης Μεταξύ Πολλαπλών Συσκευών.

Περίληψη

Παρόλο που η ψηφιακή διαφήμιση τροφοδοτεί μεγάλο μέρος του σημερινού δωρεάν Παγκόσμιου Ιστού, συνήθως λειτουργεί εις βάρος της ιδιωτικότητας των χρηστών του, λόγω της συνεχούς παρακολούθησης και διαρροής των προσωπικών δεδομένων. Προκειμένου να μεγιστοποιήσουν την ακρίβεια και αποτελεσματικότητα των διαδικτυακών διαφημίσεων, οι φορείς διαδικτυακής διαφήμισης έχουν δημιουργήσει νέες προηγμένες μεθόδους για την παρακολούθηση της περιήγησης των χρηστών σε πολλαπλές συσκευές και την προώθηση στοχευμένων διαφημίσεων στην πιο κατάλληλη οθόνη. Δυστυχώς, η τεχνική αυτή οδηγεί σε μεγαλύτερες ανησυχίες σχετικά με την ιδιωτικότητα των χρηστών, οι οποίες δεν έχουν μελετηθεί μέχρι στιγμής εκτεταμένα.

Σε αυτή την μεταπτυχιακή εργασία προτείνουμε μια νέα μεθοδολογία για την μέτρηση της τεχνικής παρακολούθησης πολλαπλών συσκευών και τη διερεύνηση των παραγόντων που επηρεάζουν την απόδοσή της, με έναν επαναλαμβανόμενο και συστηματικό τρόπο. Αυτή η νέα μεθοδολογία βασίζεται στην προσομοίωση ρεαλιστικής δραστηριότητας περιήγησης χρηστών από διαφορετικές συσκευές, που έχει ως αποτέλεσμα την συλλογή, εντοπισμό και ταξινόμηση στοχευμένων διαφημίσεων. Υλοποιώντας τη μεθοδολογία αυτή σε ένα νέο εργαλείο αυτοματοποιημένων μετρήσεων και ανάλυσης, έχουμε την δυνατότητα να πειραματιστούμε με πολλές παράλληλες συσκευές, διαφορετικούς τύπους χρηστών, και με πολλαπλά διαφορετικά πειραματικά σενάρια. Χρησιμοποιώντας το εργαλείο μας, εκτελούμε διάφορα κρίσιμα πειράματα και είμαστε σε θέση όχι μόνο να ανιχνεύσουμε και να μετρήσουμε την τεχνική αυτή με μέση ακρίβεια 78-96%, αλλά και να παρέχουμε σημαντικά αποτελέσματα σχετικά με τη συμπεριφορά των οντοτήτων που εμπλέκονται σε τεχνικές παρακολούθησης στο διαδίκτυο, αναλύοντας παράλληλα τις άμεσες επιπτώσεις στην ιδιωτικότητα των χρηστών.

Στην πραγματικότητα, η εύκολη επεκτασιμότητα της μεθοδολογίας, μας επιτρέπει να διερευνήσουμε σε βάθος αυτήν την τεχνική, και να προτείνουμε νέες επεκτάσεις για την μελέτη του οικοσυστήματος της ψηφιακής διαφήμισης. Τα ευρήματά μας μπορούν να είναι χρήσιμα τόσο για την ευαισθητοποίηση των χρηστών, όσο και για την ενίσχυση της διαφάνειας των πρακτικών παρακολούθησης που χρησιμοποιούν οι διαφορετικές οντότητες στο διαδίκτυο.

Ευχαριστίες

Πρώτα απ 'ολα, θα ήθελα να ευχαριστήσω τον επόπτη μου, καθηγητή Ευάγγελο Μαρκάτο, για την πολύτιμη καθοδήγησή του και όλες τις εποικοδομητικές συνομιλίες που είχαμε. Επίσης θέλω να εκφράσω τη βαθύτατη ευγνωμοσύνη μου στον συνεπιβλέποντα μου, Δρ. Σωτήρη Ιωαννίδη, για την την ευκαιρία να δουλέψουμε σε διαφορετικούς και ενδιαφέροντες τομείς, τα τελευταία χρόνια. Η υποστήριξη και οι συμβουλές του συνέβαλαν σημαντικά στην ακαδημαϊκή μου δραστηριότητα και την ανάπτυξη των τεχνικών εφοδίων μου. Επιπλέον, αισθάνομαι ευγνώμων στον καθημερινό μου συνεργάτη και συνεπιβλέποντα Δρ.Νικόλαο Κουρτέλλη, για την καθοδήγησή του κατά τη διάρκεια των πρώτων βημάτων του ακαδημαϊκού μου ταξιδιού και για την συμμετοχή του σε αυτή την εργασία. Οι θερμότερες σκέψεις μου πηγαίνουν επίσης στον Παναγιώτη Ηλία για τη συμβολή, την υποστήριξή και την φιλία του σε όλη τη διάρκεια της μεταπτυχιακής εργασίας.

Θα ήθελα επίσης να ευχαριστήσω όλα τα αγόρια (και φυσικά τα κορίτσια) του Εργαστηρίου Κατανεμημένων Υπολογιστικών Συστημάτων: Γιώργος Χρήστου, Δημήτρης Ντεγιάνης, Ηλίας Παπαδόπουλος, Ραφαήλ Τσίρμπας, Εύα Παπαδογιαννάκη, Μιχάλης Διαμαντάρης, Μιχάλης Παχυλάκης, Δημήτρης Καρνίκης, Γιάννης Γιακουμάκης, Μάνος Χατζημπύρος, Μάνος Αθανάτος, Δέσποινα Κοπανάκη, τον απόλυτο Χρήστο Παπαχρήστο, και όλα τα υπόλοιπα μέλη για τις συμβουλές τους και την φιλία τους.

Αυτή η μεταπτυχιακή εργασία δεν θα είχε ολοκληρωθεί χωρίς τη μακροχρόνια υποστήριξη των πιο κοντινών φίλων μου, Γιάννη, Οδυσσέα και Ολυμπίας και χωρίς την ανεκτίμητη φροντίδα και αγάπη της μητέρας και της οικογένειάς μου, που δεν σταμάτησαν λεπτό να με υποστηρίζουν.

Τέλος, αφιερώνω αυτή τη δουλειά σε Αυτήν, που έφερε ξανά το φώς στη ζωή μου.

Acknowledgements

First of all, I would like to thank my supervisor, Professor Evangelos Markatos, for his valuable guidance and all the constructive conversations we had. I also want to express my deepest gratitude to my advisor, Dr. Sotiris Ioannidis, for giving me the opportunity to work on so many different, challenging and interesting projects, over the past years. His support and advice greatly contributed to my academic and technical growth. Moreover, I feel thankful to my daily advisor Dr. Nicolas Kourtellis, for his guidance during my first steps of the academic journey and for setting the foundations of this work. My warmest regards also goes to Panagiotis Ilia for his contribution and support on this work.

I would also like to thank all the guys (and of course girls) of the the Distributed Computing Systems Laboratory: Giorgos Christou, Dimitris Deyannis, Elias Papadopoulos, Rafail Tsirmpas, Eva Papadogianaki, Michalis Diamantaris, Michalis Pachilakis, Dimitris Karnikis, Giannis Giakoumakis, Manos Chatzimpiros, Manos Athanatos, Despina Kopanaki, the o mighty Christos Papachristos and all the other present and past members, for their friendship, advice and commitment.

This work would not be achieved without the long-standing support of my closest friends Giannis, Odysseas and Olympia, and without the invaluable caring and love of my mother and my family.

Finally, this work is dedicated to Her, for bringing all thy light in my life.

Don't bend;
don't water it down;.
don't try to make it logical;
don't edit your own soul according to the fashion.
Rather, follow your most intense obsessions mercilessly.

-Franz Kafka

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Online advertising, one of the most important driving force of today's economy, shapes the socio-economic and technological landscapes with the provision of new online services and applications. It continuously grows in an unprecedented rate, to the point that it has already outperformed other more traditional ways of reaching out to the people and promoting products and services. As reported in [49], digital ad spending in 2017 has reached $209 billion worldwide and for the first time surpassed spending for TV-based advertising. The eminence of online advertising is that it can be easily tailored to the audience, and become personalized to each particular user according to her needs and interests.

To make ads more relevant, the ad-ecosystem employs various privacy-intrusive techniques to track users. There is a plurality of 3rd-party entities, created and supported by the ad-ecosystem's infrastructure, whose aim is to collect sensitive personal information, and finally use them to deliver different types of ads, from contextual to behavioral and retargeted ads. Until recently, ad-companies targeted users in regards to the activity presented in one specific device. However, as users possess multiple devices [15, 17], advertisers started moving towards more advanced practices that are specifically designed to track and target them across their devices. These advances indicate a radical transformation of the ad-landscape from *device-centric* to *user-centric*. In this new paradigm, advertisers try to identify which devices belong to the same user(i.e., a smartphone, a tablet, a laptop,etc.) and target users across them. Figure 1.1 illustrates a typical cross-device tracking (CDT) scenario, where a mobile user is targeted with relevant ads across her different devices, due to the behavior she exhibited to the ad-ecosystem from her mobile device.

According to a recent FTC Staff Report [7], CDT can be deterministic or probabilistic, and companies that engage in such practices typically use a mixture of both techniques. Deterministic tracking, which utilizes 1st-party login services that require user authentication(e.g., Facebook, Twitter, Gmail), can identify the user across multiple devices with certainty. These 1st-party services often share information (e.g., a unique identifier) to 3rd-parties, enabling them to perform

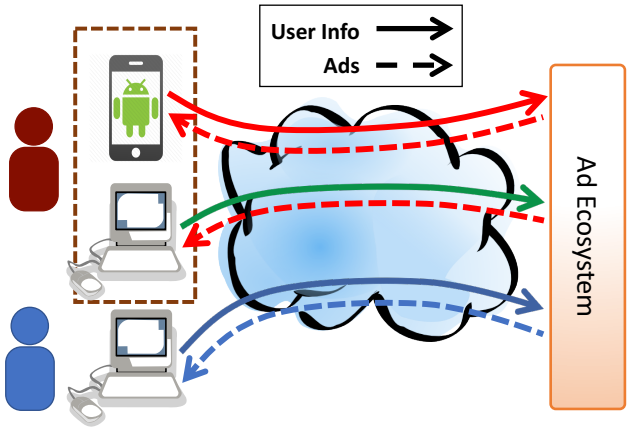

Figure 1.1: High level representation of cross-device tracking.

more effective tracking. Alternatively, in the case of probabilistic CDT, there are no shared identifiers between devices, and 3rd-parties try to identify which devices belong to the same user by considering network access data, common patterns in browsing history and behavior,geolocation metadata, etc.

In either case, the implications for user privacy are severe: ad-companies are capable of tracking individual users across all their digital space and screens, and use such information in a non-transparent fashion. In fact, in a Web that is constantly becoming more complex, there is little to no transparency on behalf of ad-companies regarding their tracking and targeting practices. Users are typically unaware of such techniques, and what's more, it is inherently difficult to measure and expose probabilistic CDT in a systematic way, as it is heavily based on user's activity. Therefore, recent privacy regulations (e.g., EU's GDPR [43]) will not be easy to enforce in such cases.

The main problem in detecting and measuring cross-device tracking lies in distinguishing which ads are presented to the user because of her behavior on that specific device (targeting or retargeting), and which ads are presented because of her activity on a different device (i.e., retargeting based on CDT). Apart from some empirical evidence about the existence of CDT, there is a limited number of studies investigating it. In one such work, Brookman et al. [29] examined 100 popular websites to determine which of them disclose data to trackers, and which pieces of data can be possibly used for the purpose of cross-device tracking. In the most closely related work, Zimmeck et al. [82], designed an algorithm that, given logs of users' browsing activity, correlates mobile and desktop devices into pairs by considering devices' browsing history and IP addresses. While this approach shows that correlation of devices is possible when such data are available, it does not provide an approach for detecting and measuring it.

In effect, our work takes the first and crucial step in understanding the inner workings of the CDT ecosystem. To the best of our knowledge, we are the first to propose a novel methodology for investigating probabilistic CDT in an automated

and systematic way, and measuring various parameters that affect its performance on the Web.

The methodology proposed in this thesis is designed based on the following idea: if cross-device tracking actually exists, and if trackers that employ such techniques (i.e., CDT-trackers) manage to successfully correlate the user's devices, it could be possible to detect it by identifying cross-device targeted behavioral ads (i.e., ads that are delivered on one device, but have been triggered because of the user's browsing behavior on a different device). In an effort to make trackers correlate the different devices of the end-user, and serve cross-device targeted ads, we employ artificially created users(dubbed as *personas*) with specific interests to emulate realistic browsing activity across the user devices. Furthermore, we built a novel framework that materializes our methodology in order to collect, categorize and analyze all the ads delivered to the different user devices, and evaluate with simple and advanced statistical methods, the potential existence of CDT.

Through a variety of novel experiments, we are able to measure CDT with 78-96% accuracy. Specifically, in the simplest experiment, where the user exhibits significant browsing activity mainly from the mobile device, we achieve average accuracy of 78% for 10 different emulated behavioral profiles. When the user exhibits significant browsing activity from both devices (mobile and desktop), with a matching behavioral profile, we observe CDT with an average accuracy of 83%. Finally, in the case of visiting specifically chosen websites that employ multiple known CDT-trackers, we observe an average detection accuracy of 96%. We also find that browsing in incognito can reduce the effect of CDT, but does not eliminate it, as trackers can perform device matching based only on the current browsing session of the user, and not all her browsing history.

1.1 Contributions

To summarize, the main contributions of this master thesis are:

- Design a novel methodology for detecting CDT by triggering behavioral cross-device targeted ads on one user device, according to specifically-crafted emulated browsing behaviors (personas), and then detecting those ads when delivered on a different device of the same user.
- The implementation of this novel methodology into CoDeT, a new and practical framework for CDT measurements. CoDeT has been designed to provide scalability for fast deployment of multiple parallel device instances, to support various configurations and experimental setups, and to be extensible for web tools and plugins available in the future.
- Conduct a set of experiments for measuring the potential existence of CDT in different types of emulated users and behaviors, with an average accuracy of 78-96%, and investigating the various factors that affect its performance under different classes of experiments.

1.2 Thesis organization

In § 2.1 we provide the necessary terminology to understand the technical contributions of our work, and in parallel we present various mechanisms and technologies proposed in related works. In § 3 an overview of the design of our methodology is provided, while in § 4 we introduce the technical parts and the implementation of the methodology into a system. In § 5 we evaluate our platform by measuring CDT under different experimental setups, and finally in § 6 and § 7 we further analyze the impact of CDT for the user's privacy, and give additional directions for future investigation.

Chapter 2

Background & Related Work

2.1 Personalized Targeted Advertising

As the purpose of online advertising is to increase market share, the advertising industry continuously develops new mechanisms to deliver more effective and highly targeted ads. These mechanisms involve the delivery of different types of ads: contextual, behavioral, targeted and retargeted. Contextual advertising refers to the delivery of ads relevant to the content of the publishing website. With regards to the effectiveness of the contextual advertisement, Chun et al. [36] found that it enhances brand recognition and that users tend to have favourable attitudes towards it. In one of the first works in this area, Broder et al. [28] proposed an approach for classifying ads and web pages into a broad taxonomy of topics, and then matching them with semantically relevant ads. Joshi et al.[48] moved ahead and proposed extending contextual advertising with behavioral information of the visitors, in order to make ads more relevant to each user.

A large body of work also investigates targeted behavioral advertising with regards to different levels of personalization, based on the type of information that is used to target the user [26, 22, 78], and its effectiveness [80, 44, 57, 14]. Interestingly, Aguirre et al. [22] found that, while highly personalized ads are more relevant to users, they increase users' sense of vulnerability. In another study, Dolin et al. [39] measured users' comfort regarding personalized advertisement. In a different direction of investigation, Carrascosa et al. [32] developed a methodology that employs artificially-created behavioral profiles (i.e., personas) for detecting behavioral targeted advertising at scale. Their approach could distinguish interest-based targeting from other forms of advertising such as retargeting. An extensive review of the literature about behavioral advertising can be found in [27].

2.2 Leakage of Personal Information

In order to serve highly targeted ads, advertisers employ various, often questionable and privacy intrusive, techniques for collecting and inferring users' personal information. They typically employ techniques,both deterministic and non-deterministic, for tracking user visits across different websites, which allow them to reconstruct parts of the users' online activity. Numerous works investigate the various approaches employed by trackers, and focus on protecting users' privacy.

In a recent work, Papadopoulos et al. [70] developed a methodology that enables users to estimate the actual price advertisers pay for serving them ads. The range of these prices can indicate which personal information of the user is exposed to the advertiser and the sensitivity of this information. Liu et al. [58] proposed *AdReveal*, a tool for characterizing ads, and found that advertisers frequently target users based on their interests and browsing behavior. Lecuyer et al. [54] proposed *XRay*, a data tracking system that allows users to identify which data is being used for targeting, by comparing outputs from different accounts. In another work, they propose *Sunlight* [55], a system that employs methodologies from statistics and machine learning to detect targeting at large scale with high statistical confidence.

Bashir et al. [25] developed a methodology that detects information flows between ad-exchanges. This approach leverages retargeted ads, in order to detect when ad-exchanges share the user's information between them, for tracking and retargeting the user. Datta et al. [38] developed *AdFisher*, a tool that explores causal connections between users' browsing activities, their ad settings and the ads they receive, and found cases of discriminatory ads. This tool uses machine learning to determine, based on the ads received, if the user belongs to a group of users that exhibit a specific browsing behavior i.e., visited specific websites that affected their behavioral profile. Castelluccia et al. [33] showed that targeted ads contain information that enable reconstruction of users' behavioral profiles, and that user's personal information can be revealed to any party that has access the ads received by the user.

In order to enable ad-targeting without compromising user privacy, Toubiana et al. [77] and Guha et al. [47] proposed *Adnostic* and *Privad*, respectively. These two approaches try to protect users' privacy by keeping user profiles on the client-side and thus, hiding user activities and interests from the ad-ecosystem. Furthermore, in an attempt to provide a better alternative, Parra-Arnau et al. [71], proposes a tool that allows users to control which information can be used for the purpose of advertising.

Furthermore, many works investigate privacy leakage, specifically, in mobile devices and the different factors influencing mobile advertising [76, 46, 63]. A recent study by Papadopoulos et al. [69] compared privacy leakage when visiting mobile websites and using mobile apps. Meng et al. [63] studied the accuracy of personalized ads served by mobile applications based on the information collected by the ad-networks. Also, Razaghpanah et al. [73] developed a technique that

detects 3d-party advertising and tracking services in the mobile ecosystem and uncovers unknown relationships between these services.

2.3 Web Tracking

As mentioned previously, various techniques are employed for tracking and correlating users' activities across different websites. Many works investigated stateful tracking techniques [74, 67, 42, 81, 56], and also stateless techniques such as browser fingerprinting [41, 19, 18, 66, 65, 68]. One of the first studies about tracking [62], investigated which information is collected by third parties and how users can be identified. Roesner et al. [74] measured the prevalence of trackers and different tracking behaviors in the web.

Olejnik et al. [67] investigated "cookie syncing", a technique that enables third parties to have a more completed view on the users' browsing history by synchronizing their cookies. Acar et al. [18] investigated the prevalence of "evercookies" and the effects of cookie respawning in combination with cookie syncing. Englehardt and Narayanan [42] conducted a large scale measurement study to quantify stateful and stateless tracking in the web, and cookie syncing, while Lerner et al. [56] conducted a longitudinal measurement study of third party tracking behaviors and found that tracking has increased in prevalence and complexity over time. and also that the most popular trackers increased appearing in popular websites.

With regards to stateless tracking, Nikiforakis et al. [66] investigated various fingerprinting techniques employed by popular trackers and measured the adoption of fingerprinting in the web. Acar et al. [19] proposed *FPDetective*, a framework to detect fingerprinting by identifying and analyzing specific events such as the loading of fonts, or accessing specific browser properties. In another work, Nikiforakis et al. [65] proposed *PriVaricator*, a tool that employs randomization to make fingerprints non-deterministic, in order to make it harder for trackers to link user fingerprints across websites. Also, in a recent work, Cao et al. [31] proposed a fingerprinting technique that utilizes OS and hardware level features, for enabling user tracking not only within a single browser, but also across different browsers on the same machine.

2.4 Cross-Device Tracking

A few recent works investigate cross-device tracking that is implemented based on technologies such as ultrasound and Bluetooth, and measure the prevalence of these approaches [61, 24, 52]. As in this work we focus on web based cross-device tracking, our work is complementary to works that investigate such technologies.

A work by Brookman et al. [29], one of the few that investigate CDT on the web, provides some initial insights about the prevalence of trackers. This work examines 100 popular websites in order to determine which of them disclose data to trackers,

identifies which websites contain trackers known to employ CDT techniques, and also investigates if users are aware of these techniques.

During the Drawbridge Cross-Device Connection competition of the ICDM 2015 conference [8], the participants were provided with a dataset [5] that contained information about some users' devices, cookies, IP addresses and also browsing activity, and were challenged to match cookies with devices and users. This resulted in a number of short papers [23, 30, 50, 51, 53, 75, 79] that described different types of machine learning approaches applied during the competition for matching devices and cookies. Some of the proposed methods achieved accuracy greater than 90%, and seen from a different point compared to our study, showed that users' devices can be potentially correlated if enough information is available. In addition, Funkhouser et al. [45] proposed a Bayesian similarity algorithm based on device characteristics and identifiers, that correlates pairs of devices with accuracy higher than 90%. This algorithm was evaluated on a dataset that contains 700 million devices along with their metadata, and outperformed other traditional unsupervised learning approaches.

Zimmeck et al. [82] conducted an initial small-scale exploratory study on CDT based on the observation of cross-device targeted ads in two "paired" devices (mobile and desktop) over the course of two months. Following this exploration, they collected the browsing history of 126 users, from which 107 have provided data from both their desktop and mobile device, and designed an algorithm that estimates similarities and correlates the devices into pairs. This approach, which is based on IP addresses and browsing history, and achieves high matching rates, shows that users' network information and browsing history can be used for pairing user devices, and thus potentially for CDT.

Overall, our work builds on these early studies, as well as past studies on web tracking on different platforms. Research around CDT is still very limited;only [82, 29] initially studied some of its aspects, but without proving its actual existence or providing a methodology for detecting it. Consequently, to fill-in this gap, we propose the first of its kind methodology, and implement a novel framework, that enables systematic investigation and measurement of probabilistic CDT.

Chapter 3

Methodology to Measure CDT

The main objective of this work is to provide a methodology for investigating cross-device tracking and its mechanics, as employed by multiple ad-ecosystem entities. In particular, we aim to design and evaluate a concrete methodology for detecting and measuring such tracking activity, as well as identifying the dominant factors that affect its performance.

Our methodology emulates realistic browsing activity of end-users (with specific interests) across different devices, and collects and analyzes all types of ads delivered to these devices. Finally, it compares those ads with baseline/controlled browsing activity to establish if cross-device tracking is present or not, at what level, and for which types of user interests.
The design objectives of this methodology are the following:

- Ability to statistically distinguish and detect probabilistic CDT in a systematic and repeatable fashion.
- Scalability, for fast deployment of multiple parallel device instances, for increased data collection.
- Distributed and decentralized, so that device instances can be launched in different geographic locations, for diversity in ad-markets and participating CDT entities.
- Support the investigation of cross-device tracking in both directions, i.e., mobile $\rightarrow$ desktop, and desktop $\rightarrow$ mobile.
- Support short and long-term experiments, for data collection in ad-hoc fashion or historically through time.
- Extensibility through modular design, so that new methods available in the future can be easily deployed and tested.

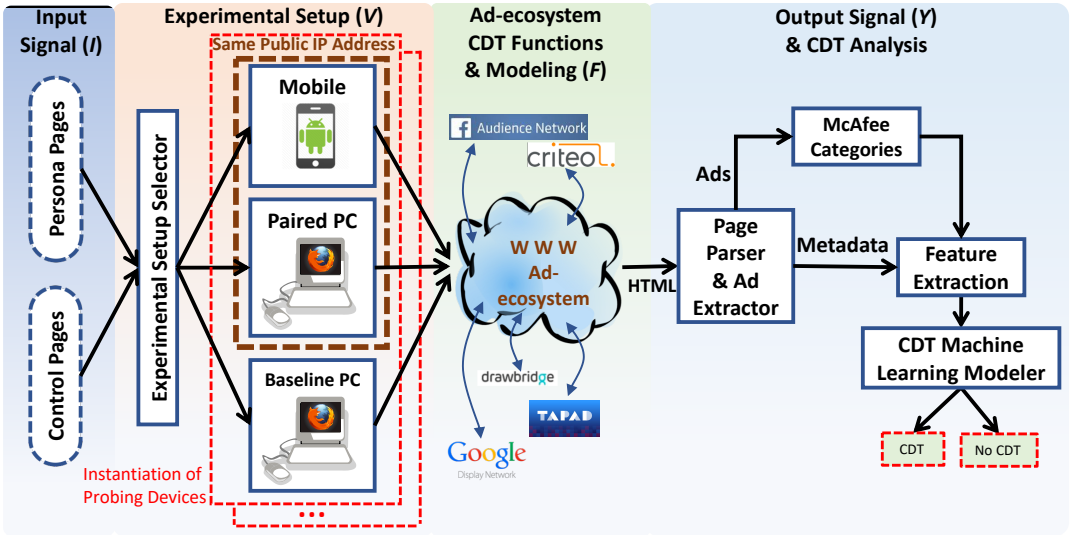

Figure 3.1: High-level representation of methodology design principles and units.

3.1 Design Principle

In general, we consider the cross-device tracking performed by the ad-ecosystem as a very complex process, with multiple parties involved, and a non-trivial task to investigate, study and understand. To that end, it is inherently difficult to identify privacy leakage due to cross-device tracking and mitigate its dangers. To infer its internal mechanics, we rely on probing the ecosystem with consistent and repeatable inputs ($\mathcal{I}$), under specific experimental settings and parameters, ($\mathcal{V}$), allow the ad-ecosystem to process and use this input via transformations and modeling ($\mathcal{F}$), and produce outputs that we can measure on the receiving end ($\mathcal{Y}$):

$$(\mathcal{I}, \mathcal{V}) \xrightarrow{\mathcal{F}} \mathcal{Y}$$

In this expression, the unknown $\mathcal{F}$ is the probabilistic modeling performed by CDT entities, which allows them to track users across their devices, regardless if these users consented to this monitoring or not.

Following this design principle, our methodology allows us to push realistic input signals to the ad-ecosystem via website visits, and measure the ecosystem's output through the delivered ads, to demonstrate if $\mathcal{F}$ enabled the ad-ecosystem to perform probabilistic cross-device tracking. The input and output can be from and to the same, or different device. Moreover, given a set of repetitions for specific experimental setups, this design allows for systematic and repeatable CDT measurements.

3.2 Methodology Challenges & Considerations

Based on this guiding design principle, an overview of our methodology is illustrated in Figure 3.1. Next, we summarize its basic assumptions and design considerations.

No 1st-party logins. Many users utilize popular online services that leak users' identifiers to 3rd-parties, making it easier to track them across different devices (since they can be identified with certainty). In our methodology, since we focus on the investigation of probabilistic CDT, we assume that the emulated user does not visit or log into any 1st-party service that employs deterministic CDT and thus, there is no common identifier (e.g., email address, OSN UID) shared between the user's devices.

Devices, IP addresses and Browsing. The approach we follow is based on triggering and identifying behavioral cross-device targeted ads, and specifically ads that appear on one of the user's devices, but have been triggered by the user's activity on a different device. For this triggering to be facilitated, the ad-ecosystem must be provided with events revealing that these two devices belong to the same user. Zimmeck et al. [82] suggest that in many cases, the devices' IP address is adequate for matching devices that belong to the same user. Relevant industrial teams [60, 21] claim that more signals can be used, such as the location of devices, browsing, etc. In fact, CDT-entities typically utilize such network-level information [35] to boost the accuracy of their methods.

Following these observations, our methodology requires a minimum of three different devices (as shown in Figure 3.1): one mobile device and two desktop computers, with two different IP addresses. We assume that two devices (i.e., the mobile and one desktop) belong to the same user, and are connected to the same network. That is, these devices have the same public IP address, are active in the same geolocation as in a typical home network, and will be considered by the ad-ecosystem as producing traffic from the same user. The second desktop (i.e., *baseline PC*), which has a different IP address, is used for receiving a different flow of ads while replicating the browsing of the user's desktop (i.e., *paired PC*). This control instance is used for establishing a baseline set of ads to compare with the ads received by the potentially paired PC.

This triplet of devices can be deployed in multiple replicas, to facilitate a faster, and large-scale data collection. Moreover, the replicas can be instantiated at different geographic locations within the same country, or even different countries, by leveraging large-scale distributed testing platforms such as PlanetLab [3], or cloud infrastructure, to probe the ad-ecosystem in different locations and collect richer and more nuanced data (ads). Such functionalities allow the methodology to be scalable, distributed and decentralized: multiple emulated devices can execute the same scripted code across different locations; then all data (ads) collected by each device can be aggregated at a centralized location for further analysis.

CDT Direction: Mobile to Desktop. In principle, our design allows the investigation of both directions of CDT. That is, users may first browse on the mobile

device, and then move to their desktop, and vice versa. However, according to a recent article [72], consumers typically use mobile devices to search for products, but make purchases on larger-screen computers. Also, ad-targeting companies such as AdBrain [20] and Criteo [37] support that the direction from mobile to desktop is more suitable for cross-device retargeting. Even though the proposed methodology allows studying both directions of CDT, in this work we focus on the mobile to desktop direction ($Mob \rightarrow PC$). In essence, the mobile device performs a specifically instructed web browsing session, i.e., *training* phase; then, the two desktop computers also perform a different type web browsing, i.e., *testing* phase, where they visit a set of pages and collect the delivered ads. The browsing performed by the two desktop devices is synchronized by means of visiting the same pages in the same order, and performing the exact same clicks.

Emulating user behavior with personas: Training Phase. To trigger CDT, we first need to demonstrate to the ad-ecosystem some activity from a user's browsing behavior ($\mathcal{I}$). In order to make the methodology systematic and repeatable, but also produce realistic browsing traffic from scripted browsers, we visit specific websites to emulate a user's behavior according to some predefined, carefully-crafted *personas*. We leverage an approach similar to Carrascosa et al. [32] to emulate browsing behavior according to specific user interests (i.e., travel and vacations, sports, shopping), and create multiple personas of different granularities, spanning from generic to more narrow categories. For each persona, our approach identifies a set of websites (*persona pages*) that have at the given time active ad-campaigns. This training activity aims to drive CDT-trackers into possible *device-pairing* between the two user's devices with high confidence.

Control pages: Testing Phase. The browsing based on a given persona can be considered as the *input* to the ad-ecosystem ($\mathcal{I}$), and the ads delivered to the involved devices as the *output* of the ad-ecosystem ($\mathcal{Y}$). To reduce any bias from possible behavioral ads delivered to specific type of websites, and following past works on this topic [32, 25], the desktops collect ads by visiting neutral websites that typically serve ads not related to their content. We refer to these neutral websites as *control pages*.

CDT Detection: Comparing Signals. Various statistical methods can be used to associate the input signal $\mathcal{I}$ of persona browsing in the mobile device, with the output signal $\mathcal{Y}$ of ads delivered to the potentially paired-desktop. For example, methods that perform similarity computation between the two signals in a given dimensionality (e.g., Jaccard, Cosine) can be of use. These methods, as well as typical statistical techniques (e.g., permutation tests) capture only one dimension of the input/output signal and thus, might not be suitable for measuring with confidence the high complexity of the CDT signal. In this case, more advanced methods can be employed, such as Machine Learning techniques (ML) for classification of the device signals as similar enough, based on features from the experimental setup ($\mathcal{V}$), and the input/output variables. In our analysis, we mainly focus on ML to compute the likelihood of the two signals being the product of CDT, as it takes into consideration this multidimensionality in the feature space.

3.3 Possible Experimentations

This methodology allows us to experiment in different ways while investigating cross-device tracking. Both persona and control pages can be used as input in either of the two types of devices (mobile or desktop). For example, the personas mechanism can be used to provide input webpages for visiting only from the mobile device, and control pages only from the desktop devices. In this case, the browsing signal for the specific persona is inputted to the ad-ecosystem from the mobile device, and the desktop devices are the recipients of the output signal. This setup, which purposely does not establish a behavioral profile on all the user's devices, aims to reveal cases of device pairing based solely on the IP address of the devices. By using two devices with the same IP address, and establishing a behavioral profile only on one of them (e.g., the mobile device), we can demonstrate the effect of pairing by detecting cross-device targeted ads on the device that has not gone under behavioral training (i.e., the desktop). This setup can be considered as providing a clearer input signal to the ad-ecosystem from the two paired devices.

Alternatively, the method can perform behavioral training on all devices, and measure the difference in the signal captured between the mobile-desktop and the mobile-baseline desktop pairs. This experimental setup aims to ease device pairing, as the devices exhibit similar browsing activity. In effect, this setup blurs the signal inputted to the ad-ecosystem, by having all devices providing similar input $\mathcal{I}$. To be able to identify cross-device tracking, such an experimentation needs to be executed for a longer period of time, to collect adequate samples for the signal comparison. Consequently, the method would compare the cumulative outcome of the user's desktop (that has the same IP address with the mobile device) with the baseline desktop PC. The former accounts for both ad retargeting (due to the desktop's browsing) and cross-device ad targeting (due to the paired mobile), while the latter only for ad retargeting (due to the baseline desktop's browsing). Finally, the selection of browsing pages to be visited by the mobile and/or desktop devices (persona and control pages) can be either generic, or specific pages that include an abundance of trackers and other third party entities specializing in CDT.

The next section details how the proposed methodology can be implemented into a real functioning system.

Chapter 4

CoDeT : A System to Measure CDT

The methodology, as already described in § 3, allows for systematic and replicable experiments that can conclusively detect evidences of cross-device tracking. A high level overview of our methodology, and its materialization by our framework CoDeT, is presented in Figure 3.1. In the following paragraphs, we describe in more detail these building blocks, and argue for various design decisions taken while implementing this methodology into the fully-fledged automated system. First, in § 4.1, we describe the process for selecting webpages to be visited by the devices (mobile and desktop) as input to the ad-ecosystem. We explain how the control pages are selected, and introduce a method for creating personas as emulated users. Second, in § 4.2, we explain the functionality of the experimental setup selector, and how mobile and desktop devices are emulated. Third, in § 4.3, we detail the methods used for parsing webpages visited by the devices to reliably extract ads and associate categories to their landing pages, and finally, in § 4.4, we present the machine learning modeler for detecting CDT within our experimental setups.

4.1 Input Signal: Personas & Control Pages

Persona Pages. A critical part of our methodology is the design and automatic building of realistic user personas. Each persona has a unique collection of visiting links, that form the set of *persona pages*. Since we do not know in advance which e-commerce sites are conducting cross-device ad-campaigns, we design a process to dynamically detect active persona pages of given interest categories. Our approach for persona generation is shown in Figure 4.1.

We use the persona categorization of Carrascosa et al. [32], for their top 50 personas, and iterate through the Google Product Taxonomy list [4], to obtain the related keywords. We do not search deeper than Level 4 (the labels below Level 4 typically correspond to very specific and rare products), of this list and we store

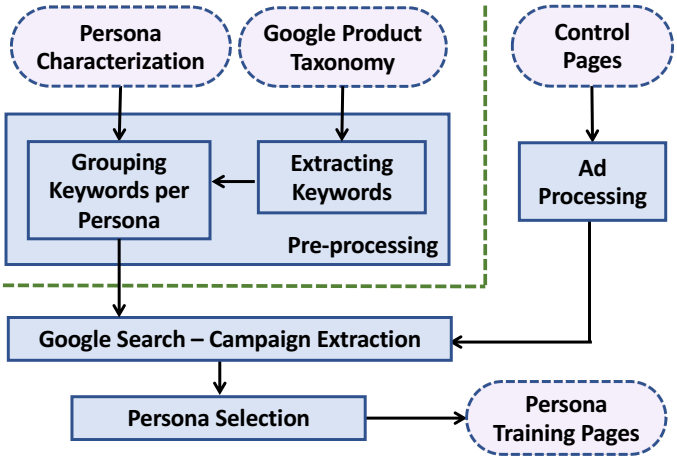

Figure 4.1: Persona design and automatic generation.

two to three keywords as a label for each persona, as we want to capture the user interests in a more general and descriptive categorization.

For capturing active ad-campaigns we use Google Search, as it reveals campaigns associated with products currently being advertised. That is, if a user searches for specific keywords (e.g., "men watches"), Google will display a set of results, including a list of sponsored links from e-commerce sites and services conducting campaigns for the terms searched. In this way, we use the keywords set for each persona, as extracted above, and transform them to search queries by appending common string patterns such us "buy, sell, offers" to create queries in a neutral fashion. This procedure is repeated until at least five (and a maximum of ten) unique domains per persona are collected. If the process fails to capture more than five unique domains, no persona is formed.

In general, our method can generate a large number of different personas, corresponding to various interests and online behaviors: from generic to specific categories. However, as the effectiveness of a persona depends on the active ad-campaigns, and also due to the computation overhead involved in creating each persona, in our experiments we only deploy the 10 personas shown in Table 4.1.

Control Pages. For retrieving the display ads from the devices, we employ a set of webpages that contain: (i) easily identifiable ad-elements and (ii) a sufficient number of ads that remain consistent through time. These pages have neutral context and therefore, do not affect the behavioral profile of the device during the visit. For most of the experiments in chapter 5 we use a set popular weather websites as control pages, similarly to the work of [32]. This set contains five webpages : { `www.accuweather.com`,`www.wunderground.com`,`www.weather.com`,`www.weather-forecast.com`,`www.metcheck.com`}. We also empirically confirmed the neutrality of ads served on this set of pages, and the lack of contextual ads. When visiting the set of control pages, our method extracts, analyzes and categorizes all the ads delivered on the active device, in order to identify those

that have been served to the user's desktop because of the browsing on the mobile device.

Table 4.1: Behavioral personas generated in this work for emulating user browsing activity.

Persona	Category - Description
1	Online Shopping - Accessories, Jewelry.
2	Online Shopping - Fashion, Beauty.
3	Online Shopping - Sports and Accessories.
4	Online Shopping - Health and Fitness.
5	Online Shopping - Pet Supplies.
6	Air Travel.
7	Online Courses and Language Resources.
8	Online Business, Marketing , Merchandising.
9	Browser Games - Online Games.
10	Hotels and Vacations.

4.2 Experimental System Setup

The experimental setup contains different types of units, connected together for replicating browsing activity on multiple devices. Typically, CDT is applied on two or more devices that belong to the same user, such as a desktop and a mobile device. Therefore, the system contains emulated instances of both types, controlled by a number of experimental parameters.

Experimental Setup Selector. As shortly described in chapter 3, we need two phases (*training* and *testing*) of browsing to different types of webpages, in order to successfully measure CDT. For that reason, we set the two browsing phases in the following way. During the training phase, the selected device visits the set of *Persona Pages* for a specific duration, referred as training time (t_{train}). The test phase is the set of visits to *control pages* for the purpose of collecting ads. During this phase, we control the duration of browsing, and we call it testing time (t_{test}). In fact, the process of training and testing is repeated several times, in order for the ad-ecosystem to be exposed repeatedly to the given signal. The experimental setup selector controls various parameters and the values selected: which device and what type will be trained, tested, the times t_{train} and t_{test}, the sequence of time slots for training and testing from the selected device, number of repetitions of this procedure, etc.

Desktop Device. The desktop devices are built on top of the web measurement framework OpenWPM [42]. This platform enables launching instances of the Firefox browser, with any set of extensions, and collects a wide range of measurements in every browsing session. It is also capable of storing the browser's specific data

(cookies, local cache, temporary files) and export a browser profile after the end of a browsing session, which can be then loaded in a future session. With these options, we can perform *stateful* experiments, as a typical user's web browser that stores all the data through time, or *stateless* experiments to emulate browsing in incognito mode offered by modern browsers. Other frameworks could also be used for automation (e.g, Selenium, PhantomJS, CasperJS) but further development is needed to support the functionalities provided by OpenWPM. We also computed via Panopticlick [9] the browser fingerprint of every device involved in CoDeT, to validate the uniqueness of the input signal to the ad-ecosystem.

Mobile Device. For the mobile device, we use the official Android Emulator [16] that allows us to create and control emulated Android devices of different vendors, OS versions etc. For the automation of browsing, we use Appium UI Automator [12], an open source automation framework, designed with native, hybrid and mobile apps. There are various applications we could use : the official UI automator by Android Studio, Robotium or Selendroid, but we chose Appium as it is compatible and easily applicable to Android Emulators and well documented. We built the mobile browsing module on top of those components to automate the visits to pages via the Browser Application. Also, our mobile browsing module provides attributes that can drive to a more realistic interaction with a website, e.g., scrolling rate, click rate, and sleep time. Similarly to the desktop device, the mobile Browser App can run either in a *stateful* or *stateless* mode. For our experiments we used a custom emulated mobile device, running Android OS version 5.4

This triplet of devices can be instantiated multiple times, depending on the computing resources available, and the experimental questions under investigation. For example, multiple instances can be executed in parallel to collect more data points on cross-device tracking faster.

4.3 Page Parser, Ad Extractor & ad-categories

Page Parser. To collect the display ads, we first need to identify specific DOM elements inside the visited webpages. This task is challenging due to the dynamic Javascript execution and the complex DOM structures generated in most webpages. For the reliable extraction of ad elements and identification of the landing pages, we follow similar methodology with Liu et. [58]. As landing pages, we refer to the destination websites that a user would be redirected to when clicking on the ads. The functionality of the Page Parser is to load the rendered HTML webpage and extract the attributes of the display ads, which also contain the landing page. In most modern websites, the display ads are embedded in nested *iFrame* tags that create deep nesting layers, containing numerous and different types of elements. However, since the ads served by our control pages are found directly inside the iFrames, the Ad Extractor described next does not have to handle such behavior.

Ad Extractor. This module is activated when the visited page is fully loaded and

no further changes occur on the content, or up to a time threshold of 60 seconds for content, and a hard timeout of 120 seconds for network responses. At first, as outlined in Algorithm 1, the module identifies all the active iFrame elements and filters out the invalid ones that have either empty content or zero dimensions. Then, it reads the *href* attributes of image and flash ads and parses the URLs, while searching for specific string patterns such as *adurl=, redirect=*, etc. These patterns are typically used by the ad-networks for encoding URLs in webpages. Finally, the module forms the list of candidate landing pages, which are then processed and analyzed to create the set of true landing pages. The ad extractor module is fully compatible with the crawlers, and does not need to perform any clicks on the ad elements (e.g., ad banners), since it extracts only the previously described data (i.e., URLs) directly from the rendered webpage. Also, the click on the ads would contribute to the known problem of ad-fraud, and impact the budget of advertisers.

Algorithm 1 Functionality of Ad Extractor.

Input: Webpage // the rendered HTML webpage
Output: LandingPages // list of candidate landing pages

```
LandingPages=∅
AlliFrames ← Collect_All_iFrames(Webpage)

for iframe in AlliFrames do
    if iframe is not_empty and visible then
        References = Collect_All_References(iframe)
        for ref in References do
            if ref contains landingpage then
                LandingPages ← Add_Reference(ref)
            end
        end
    end
end
return LandingPages
```

Ad Filter. This module processes the list of candidate URLs in order to finally obtain the true landing pages, along with their semantic category. At this phase, the platform stores only the active ad-domains, after filtering the list of landing pages with the *EasyList* [10] provided by AdBlockPlus. Similarly to previous works [25, 42] we decided to use EasyList as it is regularly updated and widely used. Other individual lists [1, 2] could also be used, or a combination of them, to enrich the Ad Filter and increase its accuracy.

Ad Categories. To associate landing pages or browsing URLs with web categories, we employ the McAfee TrustedSources database [13], which provides URLs classification based on the content of each page. This system categorized 96% of

the true landing pages of our collection into a total of 76 unique categories, by providing up to four semantic categories for each page. The remaining unclassified domains are manually classified into the categories above.

The final output of the Page Parser contains the landing pages of ads for every test phase, along with their categories. This module also stores *metadata* from the crawls such as: time and date of execution, identified ads and their categories, etc.

4.4 CDT Machine Learning Modeler

As we previously introduced, probabilistic CDT is the task of recognizing patterns of the same user across different devices, without knowing if the user is in fact the same, or any further details about the distribution or the properties of those data. This kind of task is generally suitable for investigation through Machine Learning methods, after some necessary preprocessing of the data. Previous work by Zimmeck et al. [82], as well as industry directions [60, 21] claim that probabilistic device-pairing is based on specific, well-defined signals: IP address, geolocation, type, intensity and frequency of browsing activity. In our methodology, since we control these parameters by definition, we construct the ground truth with our experimental setups. That is, we control (i) the devices used, which are potentially paired under a given IP address, (ii) the control instance of baseline desktop device, and (iii) the browsing performed with the personas.

Before applying any ML method, every instance of the input data has to be transformed into a vector of values; each position in this vector corresponds to a variable (feature). Features are different properties of the collected data: browsing activity of a user during training time, experimental setup used with the devices, time-related details of the experiment, as well as information about the collected ads, which is the output signal received from the given browsing activity.

For example, a set of features that describes the browsing activity of a user, during training time, is given in (4.4.1). Also, a set of features describing the experimental setup used with the device is given in (4.4.2), and a set of features describing the output signal received from the given browsing activity in (4.4.3).

$$\mathcal{I} = <\#\text{categories of page visited, list of categories, ...}> \tag{4.4.1}$$

$$\mathcal{V} = <\text{hour, day, type of device, repetition number, ...}> \tag{4.4.2}$$

$$\mathcal{Y} = <\#\text{ads delivered, \#categories, ad-domains, ...}> \tag{4.4.3}$$

These sets of features can be studied systematically to identify statistical association between the input and output signals. The only unknown factor here is whether the ad-ecosystem has successfully associated the devices or not, and if it has exhibited this in the output signal via ads. The feature space in this modeling contains the information related to the collected ads, and all the metadata described previously. In effect, our feature space is comprised of a union of the vectors $\mathcal{I}$, $\mathcal{V}$ and $\mathcal{Y}$, since all such features are either controlled or are measurable

by us. Thus, our data instances contain features from all three vectors (detailed report on the feature space in Table 4.2). Next, we show two examples of the two different classes:

```
Class 0 (i.e., device1=mobile, device2=nonpaired-desktop) =
<hour=11:00, day=2, repetition=3,
 persona pages categories={web services, e-shopping},
 landing  pages categories in desktop={finance, education}, ...>

Class 1 (i.e., device1=mobile, device3=paired-desktop) =
<hour=11:00, day=2, repetition=3,
 persona pages categories={web services, e-shopping},
 landing pages categories in desktop={fashion, e-shopping}, ...>
```

Notice how the activity of all three involved devices is encoded, and how the experimental setup is the same with the only difference falling in the output signal captured in the two desktops. The features provided in each example represent the state of the experiment at every moment of the experimental run. That is, what time the crawling took place for the mobile and the desktop devices, the persona categories used, ads found, etc. Note that all selected features are independent of the association between devices whether they are being paired or not.

In order for more advanced ML methods to be applied here, we transform the problem of identifying if such vectors are similar enough, into a typical Binary Classification problem. In this case, the predicted class describes the existence or absence of device-pairing, that may have occurred between the mobile device and one of the two desktops. As a *"paired"* combination, we consider the desktop device that exists under the same public IP address and (geo)location with the mobile device. The *"not paired"* combination is the mobile device and the control/baseline instance of desktop with no IP address or other relation to the mobile device. In general, these advanced statistic methods are similar, or just a lower bound of complexity to the ones employed by advertising companies.

In the next chapter, we experiment with different algorithms on the produced datasets, to obtain the best model that can decide if there has been CDT or not by the ad-ecosystem.

Table 4.2: Description of features used by datasets.

Feature Label	**Description**
Crawl_Type	The type of desktop crawl. {0 : before/test stage, 1 : after/train stage}.
Run_ID	The indexed number of run{1,4}.
Session_ID	The index of session{1,15}.
Persona_Keywords	Vector containing the keyword categories of training pages.
Mobile_Timeslot	The exact time of crawl. The 24h zone is divided into 30 minutes time slots. {0,48}(Mobile)
Desktop_Timeslot	The exact time of crawl. The 24h zone is divided into 30 minutes time slots. {0,48}(Desktop)
Mobile_Day	The day of crawl, enumerated in : {1,7}.(Mobile).
Desktop_Day	The day of crawl, enumerated in : {1,7}.(Desktop)
Mobile_Number_of_Ads	The number of ad-domains collected during a crawl.(Mobile)
Desktop_Number_of_Ads	The number of ad-domains collected during a crawl.(Desktop)
Mobile_Unique_Number_of_Ads	The number of distinct ad-domains.(Mobile)
Desktop_Unique_Number_of_Ads	The number of distinct ad-domains.(Desktop)
Mobile_Number_of_Keywords	The number of ad-categories during the crawl.(Mobile)
Desktop_Number_of_Keywords	The number of ad-categories during the crawl.(Desktop)
Mobile_Unique_Number_of_Keywords	The number of distinct ad-categories.(Mobile)
Desktop_Unique_Number_of_Keywords	The number of distinct ad-categories.(Desktop)
Mobile_Keywords	Vector containing the keyword categories for the set of landing pages per crawl for each device.(Mobile)
Desktop_Keywords	Vector containing the keyword categories for the set of landing pages per crawl for each device.(Desktop)
Mobile_Landing_Pages	Vector containing the landing pages of ads collected per crawl for each device.(Mobile)
Desktop_Landing_Pages	Vector containing the landing pages of ads collected per crawl for each device.(Desktop)

Chapter 5

Measuring CDT in the Wild

This section describes in detail the design and execution of various experiments that explore different operational settings of the framework, while measuring the appearance of CDT and its effect on the ad-ecosystem. First, in § 5.1, we provide details of the experimental setups, including type of devices used, browsing parameters, machine learning algorithms tested, and performance metrics used. In § 5.2, we present preliminary experiments as a first validation of our platform; in § 5.3 we introduce the first class of experiments designed to emulate real users' short-lived browsing behavior through different personas, and measure the existence of CDT. In § 5.4, we introduce the second experimental class, designed to emulate users' long-lived browsing behavior, where the behavioral browsing happens in multiple devices. We also study experimental setups that focus on browsing pages infused with CDT entities, in an attempt to input a stronger signal to the ad-ecosystem, and measure the improvement in CDT detection. Finally, in § 5.5, we study how functionalities available to users to avoid tracking (e.g., incognito browsing) affect CDT, across several types of personas.

5.1 Experimental Setup

Timeline of phases. Each class of experiments is executed multiple times (or runs), through parallel instantiations of the user's devices within the framework (as shown in Figure 3.1). Each experimental run is executed following a timeline of phases as illustrated in Figure 5.1. This timeline contains N sessions with three primary stages: Before, Mobile, and After. The *Before* (B_i) stage is when the two desktop devices perform a test browsing in parallel, before the mobile device is used, to establish the state of ads before the mobile device injects signal to the ad-ecosystem. The *Mobile* (M_i) stage is when the mobile device performs a train and a test browsing. This phase injects the signal from the mobile during training with a persona but also performs a subsequent test with control pages to establish the state of ads after the training. Finally, the *After* (A_i) stage is when the two desktops perform the final test browsing to establish the state of ads after the

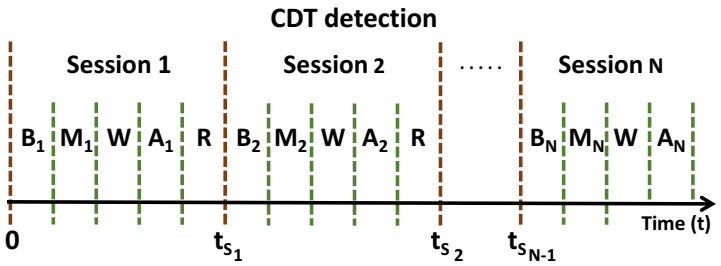

Figure 5.1: Timeline of phases for CDT measurement. M_i: mobile training; $B_i(A_i)$: testing time before (after) mobile training; W(R): wait (rest) time.

mobile training.

After extensive experimentation, we found that a minimum training time t_{train}=15 minutes and testing time t_{test}=20 minutes are sufficient for injecting a clear browsing signal over noise from the trained device to the ad-ecosystem. There is also a waiting (t_{wait}=10 minutes) and resting time (t_{rest}=5 minutes) between the stages of each session, to allow the alignment of instantiations of devices running in parallel during each session. In total, each session lasts 1.5 hours and is repeated N=15 times during a run. Through the experimental setup selector, we define the values of such variables (t_{train}, t_{test}, t_{wait}, t_{rest}, N, type of device), offering us the flexibility to measure and detect different cases of CDT.

Experimental Classes. For each of the classes of the experiments, we construct different datasets using the framework at hand and the methodology detailed in § 4, and summarized in Table 5.1. We vary the training and testing time, the number of personas used and the way their data are combined, as well as browsing functionalities such as keeping or not the state of the emulated user.

Table 5.1: Characteristics of the datasets used in each Setup (S) of experiments. S={1,2,3} are the Setups of experiments in sections 5.3 to 5.5 respectively; t_{total}: the total duration of experiment; t_{train}: the training duration; t_{test}: the testing duration; I: independent personas; C: data combined from personas; STF: stateful browser; STL: stateless browser; B: boosted CDT browsing.

S	Personas	Runs	t_{train}	t_{test}	t_{total}	Samples	Features
1a	10 (I, STF)	4	15min	20min	37 days	240	1100
1b	10 (C, STF)	-	-	-	-	2400	2201
2a	2 (I, STF)	4	480min	30min	6 days	192	600
2b	2 (C, STF)	-	-	-	-	384	750
2c	2 (I, STF, B)	4	480min	30min	6 days	192	500
2d	2 (C, STF, B)	-	-	-	-	384	576
3a	5 (I, STL)	2	15min	20min	9 days	120	450
3b	5 (C, STL)	-	-	-	-	600	880

Statistical Analysis of Device Signals. In order to analyze the similarity of signals (ads from a given category delivered in each device), we apply two different types of methods. First, to measure the similarity of distribution of ads delivered, we categorize them based on the tools described in chapter 4 and create appropriate frequency vectors, populated through time: one time vector for each device and each semantic category. We compare the signals using a two-tailed permutation test and reject the null hypothesis that the frequency of ads delivered (for a given category) come from the same distribution, if the t-test statistic leads to a p-value smaller than a significance level $\alpha < 0.05$. Given that such uni-dimensional test does not take into account the thethora of variables available in each experimental configuration, we further examine the ML methods which consider multidimensional data to decide if the ads delivered in each device are from the same distribution or not.

The ML analysis is based on three classification algorithms with different dependence on the data distribution. An easily applied classifier typically used for performance comparison with other models as a baseline, is Gaussian Naive Bayes. Logistic Regression is a well-behaved classification algorithm that can be trained as long as the classes are linearly separable. It is also robust to noise and can easily avoid overfitting by tuning its regularization and penalty parameters. Random Forest is a widely used ensemble learning learning method that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes of the individual trees. For the identification of the important variables on each experimental setup, we also use the Extra-Trees classifier, and the Gini Index metric, to select the subset of the most relevant features.

A fundamental point at the study of the performance evaluation of machine learning algorithms is the selection of the appropriate metrics. In general, pure accuracy can be used, but it's not representative for our analysis. The reason is that we want to report the most accurate estimation for the number of predicted paired devices, while at the same time we want to measure the absolute number of miss-classified samples overall. For this reason, metrics like Precision and Recall are typically used, since they quantify this type of information. The definitions of those metrics are the usual ones, i.e.:

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}$$

F_1-score, the harmonic mean of Precision and Recall, is also used to capture the trade-off between these two metrics.

On this type of analysis, there is also one more metric, measuring the dependence of the True Positive Rate (TPR) with the False Positive Rate (FPR). TPR is the same as Recall, and FPR is defined as follows:

$$FPR = \frac{FP}{FP + TN}.$$

If we plot the curve of those two rates for different operational scenarios, we get the Receiver Operating Characteristic curve (ROC). The curve itself gives useful insights about the relation between different types of errors of the classifier. If a single numeric score based on the ROC curve is needed, then the Area Under the Curve (AUC) is used, i.e., the integral of the ROC curve over the interval (0, 1), where the bigger the value, the better performance of the given classifier. Extensive description of metrics and algorithms used can be found on [64].

5.2 Platform Validation for Ad Measurements

A first set of preliminary experiments is introduced here, to demonstrate that our platform can (i) successfully identify and collect the ads delivered in multiple devices (mobile and desktops), (ii) inject browsing signal from a device, thus biasing it to have a realistic persona and (iii) lead to matching/pairing of devices, which could be due to simple or advanced targeting, retargeting ads or CDT.

First, we use a simple experimental setup: we connect three instances of desktop device and one mobile device under the same IP address. We create one persona, with the use of our component in section 4.1, in the interest of "Online Shopping-Fashion, Beauty", and following the given timeline of phases 5.1, we run this experiment for two days. Then, we perform one-dimensional statistical analysis, as introduced in section 5.1, and find that there is no similarity between the mobile with any of desktop devices (null hypothesis rejected with highest p-value=0.030), while all desktop distributions are similar to each other (null hypothesis accepted with lowest p-value=0.33). These results signify that in the level of ad-distribution there is no clear device-pairing for the given persona, and that we should consider controlling more factors to instigate it.

Consequently, we expand this experiment by also training one of the desktop devices using the same persona as with mobile. By repeating the same statistical tests, we find that the mobile and desktop with the same browsing behavior receive ads coming from the same distribution (null hypothesis accepted with lowest p-value=0.84), while the other devices show no similarity with each other or the mobile (null hypothesis rejected with highest p-value=0.008). The outcome of this experiment is one of our initial indications of device-pairing since the browsing behavior under a shared IP address can boost the signal towards advertisers, which they can use it to apply advanced targeting, either as CDT, or retargeting on each device or a mixture of such techniques. Finally, these preliminary experiments and statistical tests provide us some first evidences for the effectiveness of our framework to inject enough browsing signal from different devices under selected personas, to collect ads delivered that can be further analyzed. Next, we present more elaborate experimentations with our framework, in order to study CDT in action.

5.3 Detecting CDT in Short-lived Browsing

Independent Personas: Setup 1a. This experimental setup emulates the behavior of a user that browses frequently about some topics and interests, but in short-lived sessions in her devices. Given that most users do not frequently delete their local browsing state, this setup assumes that the users' browser keeps all state, i.e., cookies, cache, browsing history. This enables trackers to identify users more easily across their devices, as they have historical information about them. In this setup, every experimental run starts with a clean browser profile, and the cookies and also the browser session's files are stored for the whole duration of the experimental run. We use all personas of Table 4.1, and the data collection for each persona lasts ∼4 days.

We perform the same statistical analysis as in section 5.2, and find that in 4/10 personas, the mobile and paired desktop ads are similar (null hypothesis accepted with lowest p-value=0.13), while the mobile and baseline desktop ad distributions are different (null hypothesis is rejected with highest p-value=0.009). This inconsistency is reasonable since the statistical analysis is based only on one dimension, the frequency of specific types of ads appearing in the devices, which may not be enough for fully capturing the existence of device-pairing. Also, the IP address might not be a sufficient factor for the ad-ecosystem to lead to CDT. As already argued, we use more advanced multidimensional ML methods, to effectively compare the potential CDT signals captured by the two devices.

The classification results of the Random Forest algorithm,the best performing one compared to the other two, are reported in Table 5.2. (Naive Bayes and Logistic Regression perfomances are being reported in Appendix, Tables A.1 and A.2) We use AUC score as the main metric score in our analysis, since the ad-industry seems to prefer higher Precision scores over Recall, as the False Positives have greater impact on the effectiveness of ad-campaigns.[1] As shown in Table 5.2, the model achieves high AUC score for most of the personas, with a maximum value of 0.84. Specifically, the personas 2, 4 and 8 scored highest in AUC, and also in Precision and Recall, whereas persona 6 has poor performance compared to the others. These results indicate that for high scoring personas, we captured the active CDT-campaigns, while the ML model succesfully corellated the experimental variables, and classified the signals. For the personas with lower scores, there may not be active ad-campaigns during the period of the experiments, or a more intense experimentation is needed to distinguish the CDT signal.

The performance variation of CDT over time, based on the average AUC score of our personas, is shown in Figure 5.2. We observe that there is a higher standard error of measurement in the first few sessions, which is reasonable considering the diverse scoring of the individual personas, and the underfitting/overfiting behavior of the algorithm. After session 3 the average AUC score stabilizes, with an upward

[1] Tapad [6] reports: "Maintaining a low false positive rate while also having a low false negative rate and scale is optimal. This combination is a strong indicator that the Device Graph in question was neither artificially augmented nor scrubbed."

Table 5.2: Performance evaluation for Random Forest in Setups 1a and 1b. Left value in each column is the score for Class 0 (C0=*not paired desktop*); right value for Class 1 (C1=*paired desktop*).

Persona (Setup)	**Precision**		**Recall**		F_1**-Score**		**AUC**
	C0	**C1**	**C0**	**C1**	**C0**	**C1**	
1 (1a)	0.89	0.60	0.57	0.90	0.70	0.72	**0.73**
2 (1a)	0.84	0.78	0.81	0.82	0.82	0.80	**0.82**
3 (1a)	0.81	0.73	0.78	0.76	0.79	0.74	**0.76**
4 (1a)	0.87	0.78	0.87	0.78	0.87	0.78	**0.82**
5 (1a)	0.94	0.65	0.68	0.93	0.79	0.76	**0.80**
6 (1a)	0.57	0.67	0.81	0.38	0.67	0.48	**0.59**
7 (1a)	0.81	0.87	0.89	0.76	0.85	0.81	**0.81**
8 (1a)	0.86	0.85	0.89	0.81	0.87	0.83	**0.84**
9 (1a)	0.74	0.90	0.91	0.73	0.82	0.81	**0.81**
10 (1a)	0.77	0.85	0.81	0.81	0.79	0.83	**0.81**
Combined (1b)	0.77	0.84	0.81	0.84	0.82	0.84	**0.89**

trend during the last sessions, while the error rate constantly decreases to the point of reaching its minimum value between sessions 8 and 10 (i.e., ∼12-15 hours). This trend demonstrates that under this specific setup, the ad-ecosystem is able to correlate the paired devices within a few hours, without the need of extensive user browsing activity. Further browsing activity (*training*) from both paired devices, would boost this performance to higher levels, and also could probably result into faster device pairing.

In order to retrieve the variables that affect the discovery and measurement of CDT, we applied the feature importance method on the dataset of each persona, and selected the top-10 highest scoring features. For the majority of the personas (7 out of 10) the most important features were the number of ads (distinct or not) and the number of keywords in desktop. In some cases, there were also landing pages that had high scoring, but this was not consistent across all personas.

Combined Personas: Setup 1b. Here, we use all the datasets collected individually, for each persona in the previous experiment (Setup 1a), and combine them into one unified dataset. This setup emulates the realistic scenario of a user exhibiting multiple and diverse web interests, that give extra information to the ad-ecosystem about their browsing behavior. Of course, in this combined dataset there is an s in the feature space to accommodate all the domains and keywords from all personas. In fact, it contains 2021 features as it stores the vectors of landing pages and keywords, for all the different types of personas. In total, there were 890 distinct ad-domains described by keywords in 76 distinct categories.

The ads delivered in all three devices during these sessions are shown in Figure 5.3 (left). For most sessions (∼90%), the mobile device was exposed to less than five ads, since the mobile versions of websites typically deliver a smaller number of ads, designed to fit in smaller screens and devices. On the contrary, the

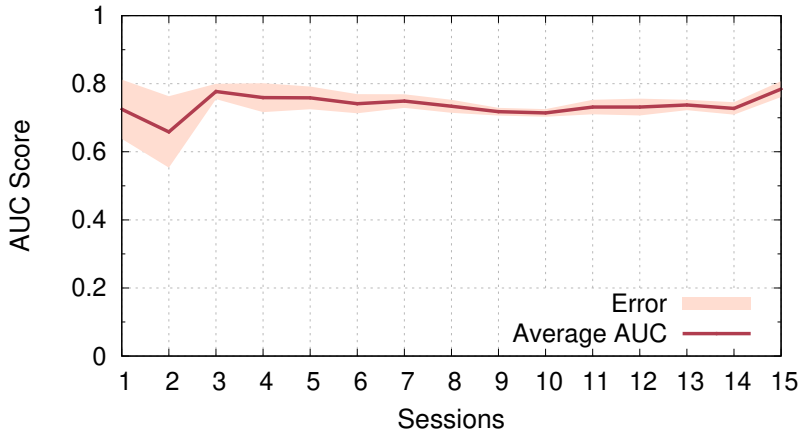

Figure 5.2: Average AUC score and standard error of measurement across the personas, combining all 4 experimental runs.

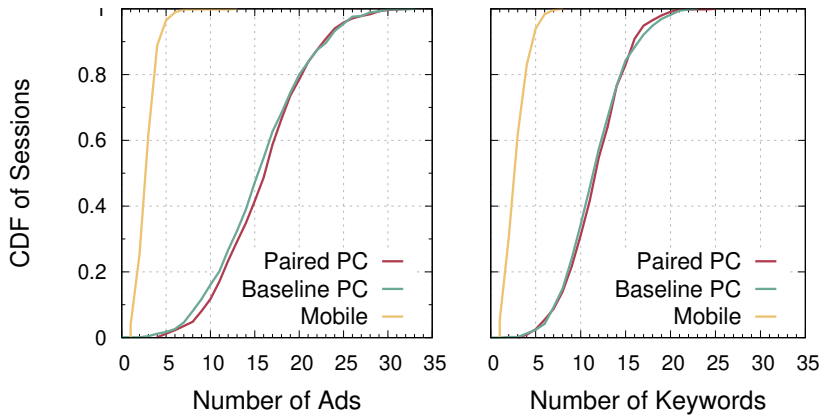

Figure 5.3: CDF of collected ads (left) and corresponding keywords of the ads (right) per crawling session for all devices.

desktop devices had a higher exposure to ads compared to the mobile device. Also, the two desktops receive a similar number of ads, on average 2 to 4 ads on every visit to the control pages. Similar observations can be made for the keywords categories of ads(Figure 5.3 (right)).

Additionaly, Figure 5.4 reports the occurrence of the top-10 keywords of the mobile and their frequency of occurrence in the other devices. The most frequent term in the mobile keywords is “Online Shopping”, since many of the personas were related to interests that involved shopping. The two desktops appear to have similar distribution for the top keywords of the mobile, and seem to be fairly different than the mobile. However, in some cases like online shopping and travel, the paired desktop turns out to be closer to the mobile. Interestingly, even though the two desktops have similar distribution for the top keywords of the mobile, and

are different than the mobile, the paired desktop appears to have some keywords which are closer to the paired mobile, hinting to the effect of device-pairing.

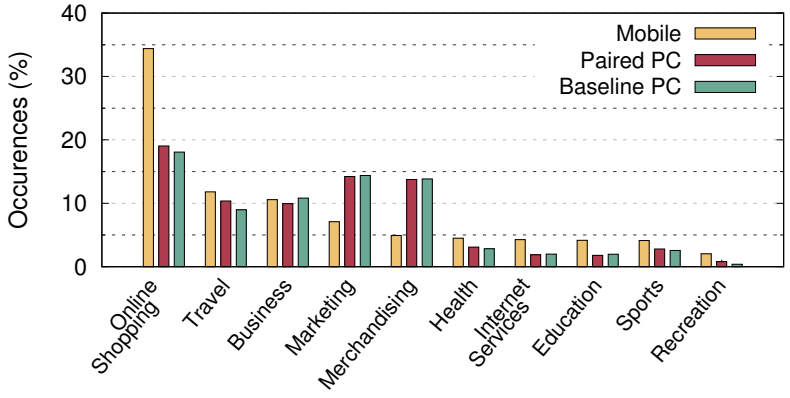

Figure 5.4: Top-10 frequent mobile keywords and their corresponding frequency of occurrence in the three devices.

In this dataset, we apply feature selection with the Extra-Trees classifier to select the most relevant features and create a more accurate predictive model. This method reduced the feature space to 984 useful features out of 2201. Next, we use the three classification algorithms and a range of hyper-parameters for each one. Also, we apply the 10-fold nested cross-validation method for selecting the best model (in terms of scoring performance) that can give us an accurate, non overly-optimistic estimation [34]. Again, the best selected model was Random Forest, with 200 estimators (trees) and 200 depth of each tree, scoring AUC=0.89 (all results in Table 5.2).

The model's performance is high in all the prior scores, which indicates that the more diverse data the advertisers collect, the easier it is for them to identify the multiple user's devices. This result is in line with Zimmeck et al. [82], who attempted a threshold-based approach for probabilistic CDT detection on real users' data, lending credence to our platform's performance.

We also compute the feature importance, shown in Figure 5.5, for the top-30 features. One third of the top features are related to crawl specific metadata, whereas about half of the top features are keyword-related. Interestingly, features such as the day and time of the experiment, and the number of received ads are important for the algorithm to make the classification of the devices. Furthermore, time-related features are indeed expected to be important as they give hints on when the browsing signal was injected to the ad-ecosystem. In addition, keyword-based features are important for the classification, revealing that since ads of similar categories get delivered in paired devices instead of non-paired, the keywords help our classifier to identify potential device pairing. These results give credence to our initial decision to experiment in a continuous fashion with regular sessions injecting browsing signal, while at the same time measuring the output signal via delivered ads.

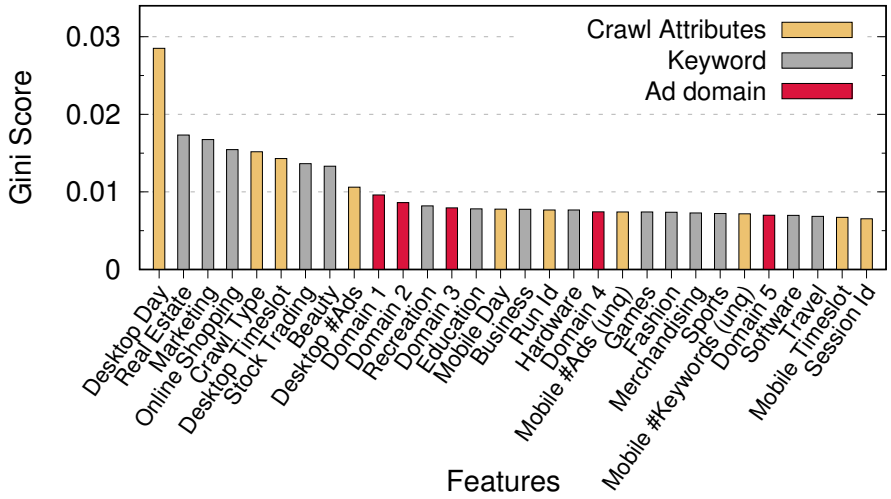

Figure 5.5: Top-30 features ranked by importance in the machine learning model.

Data Validation. We validate the representativeness of the data collected from these experiments by examining the trackers that appear in the pages visited by the personas, as well as the landing pages for each device. We use the Ghostery plugin [11] to detect them and measure their frequency of inclusion. From the trackers detected in the persona and control pages, and using the list provided by [82], 27% were found to be CDT-related, including both deterministic and probabilistic. In fact, the top-10 trackers, which may perform both types of CDT, include Google, Google Syndication, Google Analytics, Doubleclick, Facebook, YouTube, Criteo, Trivago, Advertising.com and Krux, which are in line with the top CDT trackers found in [29, 82]. Also, 14.2% of these trackers are explicitly focused on CDT, including Criteo, BlueKai, AdRoll, Cardlytics, Drawbridge, again in line with the results in [82].

Going a step further, in Figure 5.6 we analyze the trackers appearing in the top-10 most frequent landing pages (ads) for each device. The most frequent ones are Google and Facebook, while Criteo, a well-established CDT-tracker is found with a higher frequency on the paired desktop, hinting the existence of CDT in our dataset. The high appearance of Google, Facebook, and Doubleclick trackers, and the well known collaboration between them (sharing data, cookie syncing, etc.), entails that if the user accessed such webpages and leaked any identifier, the cross-device tracking would be very effective. Also, these collaborations accompanied by the variety of third parties, reveal the complexity and plurality of the ad-ecosystem, making the process of distinguishing the effectiveness of each individual CDT tracker a non-trivial task.

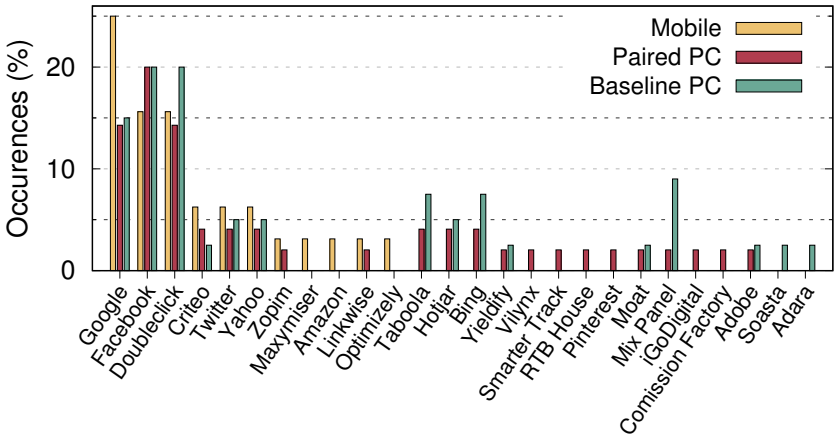

Figure 5.6: Trackers in top-10 frequent landing pages of each device set. We assume that the Google Network covers sites using AdMob, AdSense, Blogger, Google Syndication, Google Analytics, YouTube and the other tracking domains focused on different type of platforms and devices.

5.4 Detecting CDT in Long-lived Browsing

Independent Personas: Setup 2a. In this set of experiments, we allow the devices to train for a longer period of time, to emulate the scenario where a user is focused on a particular interest, and produces heavy browsing activity around it. This long-lived browsing injects a significantly higher input signal to the ad-ecosystem than the previous setups, which should make it easier to perform CDT. In order to increase the potential pairing complexity, and make it more difficult to track the user, we allow all devices (i.e., 1 mobile, 2 desktops) to train in the same way under the same persona. In effect, this setup also tests a basic countermeasure from the user's point of view, who tries to blur her browsing by injecting traffic of the same persona, from multiple devices to the ad-ecosystem. Also, while all devices are trained with the same profile, we examine here if the statistical tests and ML modeler can still detect and distinguish the CDT.

This experiment contains three different phases during each run. The mobile phase, where the mobile performs train crawls for t_{train}=480 mins, and a testing crawl for t_{test}=30 mins. In parallel with the mobile training, the two desktops perform test crawls for t_{test}=30 mins. After mobile training and testing, both desktops start continuous train and test crawls alternately for 8 hours (t_{train}=t_{test}=30 min). Due to the long time needed for conducting this experiment, we focus on two personas constructed in the following way. We use the methodology for persona creation as described in section 4.1, and focus on the currently active ad-campaigns, resulting in two personas under the interest of "Online Shopping-Accessories", and "Online Shopping-Health & Fitness" (loosely matching the personas 1 and 4 from Table 4.1). Then, we performed 4 runs of 16 hours duration each, for the two personas.

Table 5.3: Performance evaluation for Logistic Regression in total components of Setup 2. Left value in each column is the score for Class 0 (C0=not paired desktop); right value for Class 1 (C1=paired desktop).

Persona (Setup)	**Precision**		**Recall**		F_1**-Score**		**AUC**
	C0	**C1**	**C0**	**C1**	**C0**	**C1**	
1 (2a)	0.90	0.79	0.82	0.88	0.86	0.83	**0.85**
4 (2a)	0.83	0.79	0.81	0.81	0.82	0.80	**0.81**
combined(2b)	0.87	0.92	0.92	0.87	0.89	0.90	**0.89**
1 (2c)	0.87	1.0	1.0	0.88	0.93	0.93	**0.93**
4 (2c)	1.0	0.98	0.98	1.0	0.99	0.99	**0.99**
combined(2d)	1.0	0.86	0.88	1.0	0.93	0.93	**0.93**

The statistical analysis for this experiment reveals potential CDT, since we accept the null hypothesis for the distribution of ads delivered in the paired desktop and mobile (lowest p-value=0.052), and reject it in the baseline desktop and mobile (highest p-value=0.006). This consistency is interesting, since for this setup all three devices are uniformly trained with the same persona, and thus all of them collect similar ads due to retargeting. However, there is no similarity between the distributions of ads in the devices that do not share the same IP address.

To clarify this finding, we applied the aforementioned ML algorithms, and the classification results are shown in Table 5.3(all results in Appendix Table A.3). In this setup, since all the devices are uniformly trained, we did not include the keyword vector of the persona pages into the datasets, to not introduce any bias in the classifiers from repetitive features. To the point, the algorithms again detect CDT between the mobile and the paired desktop, even though all devices were exposed to similar training with the same persona. In fact, Logistic Regression performed the best across both personas, with AUC $\geq$ 0.81, and F_1-score $\geq$ 0.80 for both classes.

When computing the importance of features, the number of desktop ads and keywords and the desktop time slot are in the top-10 features overall. Based on these observations, we believe that the longer training time, allowed the ad-ecosystem to establish an accurate user profile, and perform retargeting on the paired desktop, based on the mobile's activity. The device-pairing was possible, even though there was a competing baseline desktop attempting to "scramble" the input signal to the ad-ecosystem.

Combined Personas: Setup 2b. We follow a similar approach as before (section 5.3) and combine all data collected from the Setup 2a into a unified dataset. Under this scenario, in which we mix data from both personas, the classifier again performs well, with AUC=0.89. Important features in this case are the number of ads and keywords delivered to the desktops, the time of the experiment, and number of keywords for the desktop. Once more, we observe a consistency in the high performance of the "uniform" persona (as in 5.3), which implies the fact

that the plurality of the collected user's data, produced by well-defined and intense browsing activity, increases the performance of CDT.

Boosted Browsing with CDT trackers and Independent Personas: Setup 2c. In the next set of experiments, we investigate the role of CDT trackers in the discovery and measurement of CDT. In particular, we attempt to boost the CDT signal, by visiting webpages with higher portion of CDT-type trackers. Therefore, the experimental setup and the preprocessing method remain the same as in the previous Setup 2a, but we select webpages to be visited that have active ad-campaigns and embed the most-known CDT trackers: Criteo, Tapad, Demdex, Drawbridge. We also change the set of our control pages, so that each one contains at least one CDT-tracker. News sites have a plurality of 3rd-parties compared to other types of sites [42]. Thus, for boosted browsing, we chose the set of control pages for this experiment to contain 3 weather pages and 2 news websites. Specifically the set of control pages now is parted from:{ `www.accuweather.com,www.wunderground.com,www.weather.com,www.usatoday.com,www.huffingtonpost.com`}. Since the neutrality may not be applied to all such sites, we also manually verified that the selected pages do not serve contextual ads.

Performing the same analysis as earlier, we find that mobile and paired desktop have ads coming from the same distribution (lowest p-value=0.10), and that there is no similarity between the ads delivered in the mobile and baseline desktop (highest p-value=0.007). For a clearer investigation of the importance of the CDT-trackers, we also evaluate the findings with the ML models. The evaluation results of this experiment are presented in Table 5.3.(all results in Appendix Table A.3).

For persona 1, Logistic Regression and Random Forest models perform near-optimally, with high precision of Class 1, high recall for class 0, average F_1-Score=0.93 for both classes, and AUC=0.93. For persona 4, the scores are even higher, outperforming the other setups and experiments, as all metrics for Logistic Regression scored higher than 0.98. Overall, these results indicate that we successfully biased the CDT detection, by tricking the trackers to identify the emulated user in both devices, and providing enough output signal (ads delivered) for the statistical algorithms to detect the CDT.

Boosted Browsing with CDT trackers and Combined Personas: Setup 2d. We follow a similar approach with before, and combine all data collected from the Setup 2c, into a unified dataset for Setup 2d. Under this scenario, the classifier (Logistic Regression) again performs very well, with AUC=0.93. Important features in this case are the number of ads delivered to the desktops, the time of the experiment in each desktop and the number of keywords. Interestingly, and perhaps unexpectedly, the existence of Criteo tracker in a landing page, is a feature appearing in the top-10 features. Indeed on parallel with previous outcomes regarding longer training, the interaction with CDT-entities, totally affects and biases the device pairing.

5.5 Incognito Browsing to the Rescue?

Independent Personas: Setup 3a. In this final experimental setup, we investigate the possibility for the user to apply some basic countermeasures to avoid, or at least reduce the possibility of CDT, by removing her browsing state in every new session. For this, we perform experiments where the traditional tracking mechanisms, i.e., cookies, cached session files, browsing history, etc., are disabled or removed, emulating incognito browsing. We select the first five personas from Table 4.1, which had the most active ad-campaigns, and appeared to be promising due to the online shopping domain of interest. Every desktop executed browsing in a stateless mode, while the mobile device in a stateful mode. For each of the five personas we collected data for two runs, following the timeline of phases as in Setup 1a.

The distributions between mobile vs. paired desktop, as well as mobile vs. baseline desktop, were found to be different (highest p-value=0.034). Also, none of the ML classifiers performed higher than 70% (in all metrics), and thus we could not clearly extract any significant result. More specifically, the highest AUC score for personas 1 and 2 was 0.70 with the use of the Random Forest classifier, and for personas 3 and 4 was 0.73 using the Logistic Regression classifier. The worst scoring, independent of algorithm, was recorded for persona 5, with AUC=0.57, and Precision/Recall scores under 0.50.

Combined Personas: Setup 3b. When the data from all five personas are combined, the classifier performing the best, was Logistic Regression with AUC=0.79. Overall, these results point to the semi-effectiveness of the incognito browsing to limit CDT. That is, by removing the browsing state of a user on a given device, the signal provided to the CDT entities is reduced, but not fully removed. In fact, when the data from various personas are combined, the CDT remains somewhat effective, but still eliminated compared to the other setups. Moreover, it seems that user's tracking in one or multiple devices is still possible, since the ad-networks are capable of correlating pairs of devices only based on the activity shown from a specific IP address and (geo)location shared between them.

Chapter 6

Discussion

Through extensive experiments with CoDeT, we were able to trigger CDT trackers into successfully pairing the emulated users' devices, which allowed us to verify that CDT is indeed happening, and measure its effectiveness on different user interests and browsing behaviors, independently and in combination. In fact, CDT was very prominent when the user devices were trained to browse pages of similar interests, reinforcing the behavioral signal sent to CDT entities, and specifically when the browsing activity is related to online shopping, since those types of users seem to be more targeted by the advertisers. The CDT effect was further amplified when the visited persona and control pages had embedded CDT-trackers, pushing the accuracy of detection up to 99%.

Our analysis also showed that well-known advertising companies such as Google and Facebook, which have a prominent role in the ad-ecosystem, even if they are not applying probabilistic CDT directly, can potentially impact the spread of CDT as a practice. We compared and validated our results with past works, providing support to our platform's representativeness of persona building and data collection. As a basic counter-measure for the user, we tested the effect of incognito browsing to the CDT performance. Browsing in a stateless mode showed a reduced, but not completely removed CDT effect, as incognito browsing obfuscates somewhat the signal sent to the ad-ecosystem, but not the network access information. Furthermore, when combining data from different personas, CDT was still prominent even in incognito browsing.

Indeed, our data collection was performed across relatively short time periods, in comparison with the wealth of browsing data that CDT companies have at their disposal. In fact, we anticipate that CDT companies collect data about users, devices and browsing behaviors for months or years, and even buy data from data brokers, to have the capacity of cross-device tracking and targeting users with even higher rates. To that end, we believe that high accuracies self-reported by CDT companies (e.g., Lotame: >90% [59], Drawbridge: 97.3% [40]), are totally possible.

6.1 Future work

In order to have a more thorough investigation of CDT in the future, we should deploy more paired devices running in parallel, for longer periods of time, with different device characteristics, across multiple locations around the world, using more personas, etc. In fact, since the ad-ecosystem employs various techniques for targeting different kinds of users, one major line of future work is the study of targeting sensitive user categories (e.g., gender, sexual orientation, etc.). Moreover, a future study with real users' data could be also conducted. By doing so we can finally compare the performance of our tool given emulated and real data, and gain more powerful insights about the targeting techniques in practice.

In addition, other crucial factors not described on this work, can be easily studied in the future with extensions of our framework. Those factors include CDT's intensity, lifetime, and defense mechanisms. Another significant issue that has not yet been studied, is the efficiency of anti-tracking tools, that in theory eliminates the effect of device tracking. Finally, we also plan to expand our platform and study the Deterministic Cross-Device Tracking in the same direction, since there are no other related works in this area. Further investigation on this direction is needed, so as to understand the internal mechanisms the complexity of the ad-ecosystem and its tracking paradigms.

Chapter 7

Conclusion

In this master thesis, our initial effort was focused on defining, executing and analyzing basic experimental setups which provide conclusive and statistically significant evidence for the detection and investigation of CDT. Undoubtedly, cross-device tracking has a strong impact on user privacy, but the actual extent of this tracking paradigm and its consequences to users, the community, and even to the ad-ecosystem itself, are still unknown.

Ultimately, the rising popularity of the digital world, and the radical reformation of the ad-ecosystem does not necessarily mean that customers must give up their privacy. While companies will always need data for improving the effectiveness of their marketing practices, it is important to think about how this information is collected, stored, and traded. End-users should have a choice regarding the access control policies of their data, the kind of data that are being collected , the retention periods, etc. Those facts are especially relevant nowadays with the enforcement of recent EU privacy regulations such as GDPR and ePrivacy.

This is where our platform comes into play, as it provides a concrete, scalable and extensible methodology for experimenting with different scenarios, understanding CDT's mechanics and measuring its impact. In fact, the modular and extensible design allows the community to investigate CDT in depth and propose new extensions to study the ad-ecosystem: new plugins, personas and ML techniques. To that end, our design constitutes CoDeT into an enhanced transparency tool that reveals potentially illegal biases or discrimination from the ad-ecosystem.

Appendix A

Performance Evaluation for Setups 1 and 2

Table A.1: Performance evaluation Metrics for Naive Bayes - Setup 1a

Persona	Precision		Recall		F_1-Score		AUC
	C0	C1	C0	C1	C0	C1	
1	0.79	0.83	0.83	0.80	0.81	0.80	**0.81**
2	0.69	0.81	0.75	0.75	0.71	0.75	**0.75**
3	0.72	0.69	0.82	0.55	0.77	0.61	**0.75**
4	0.75	0.85	0.88	0.71	0.81	0.77	**0.79**
5	0.60	0.69	0.71	0.58	0.65	0.63	**0.64**
6	0.70	0.67	0.84	0.46	0.76	0.55	**0.65**
7	0.78	0.63	0.41	0.89	0.64	0.74	**0.70**
8	0.67	0.74	0.67	0.74	0.64	0.74	**0.70**
9	0.64	0.80	0.88	0.50	0.74	0.62	**0.68**
10	0.70	0.79	0.70	0.79	0.70	0.79	**0.74**

Table A.2: Performance evaluation Metrics for Logistic Regression - Setup 1a

Persona	**Precision**		**Recall**		F_1**-Score**		**AUC**
	C0	**C1**	**C0**	**C1**	**C0**	**C1**	
1	0.76	0.96	0.95	0.79	0.84	0.86	**0.86**
2	0.88	0.88	0.88	0.88	0.88	0.88	**0.87**
3	0.56	0.62	0.65	0.52	0.60	0.57	**0.58**
4	0.74	0.71	0.77	0.68	0.75	0.70	**0.72**
5	0.65	0.56	0.65	0.56	0.65	0.56	**0.60**
6	0.63	0.54	0.67	0.50	0.65	0.52	**0.58**
7	0.67	0.67	0.59	0.74	0.62	0.70	**0.66**
8	0.93	0.76	0.62	0.96	0.74	0.85	**0.79**
9	0.90	0.74	0.84	0.82	0.87	0.78	**0.83**
10	0.79	0.89	0.92	0.74	0.85	0.81	**0.82**

Table A.3: Evaluation Metrics for each Persona of experimental Setups 2a/2c. NB is acronym for Naive Bayes algorithm, LR for Logistic Regression, and RF for Random Forest.

Pers.	**Setup**	**Alg.**	**Precision**		**Recall**		F_1**-Score**		**AUC**
			C0	**C1**	**C0**	**C1**	**C0**	**C1**	
1	**2a**	NB	0.83	0.88	0.90	0.78	0.86	0.82	**0.84**
		LR	0.90	0.79	0.82	0.88	0.86	0.83	**0.85**
		RF	0.93	0.84	0.76	0.95	0.84	0.89	**0.85**
	2c	NB	0.90	0.92	0.9	0.92	0.90	0.92	**0.90**
		LR	0.87	1.0	1.0	0.88	0.93	0.93	**0.93**
		RF	0.87	1.0	1.0	0.88	0.93	0.93	**0.93**
4	**2a**	NB	0.64	0.80	0.81	0.62	0.71	0.70	**0.71**
		LR	0.83	0.79	0.81	0.81	0.82	0.80	**0.81**
		RF	0.83	0.72	0.75	0.81	0.79	0.76	**0.78**
	2c	NB	0.98	0.98	0.98	0.98	0.98	0.98	**0.97**
		LR	1.0	0.98	0.98	1.0	0.99	0.99	**0.99**
		RF	0.90	0.97	0.98	0.86	0.94	0.92	**0.92**

Bibliography

[1] https://disconnect.me/trackerprotection.

[2] https://easylist.to/easylist/easyprivacy.txt.

[3] Planetlab: An open platform for developing, deploying, and accessing planetary-scale services. https://www.planet-lab.org/.

[4] https://www.google.com/basepages/producttype/taxonomy.en-US.txt, 2015.

[5] ICDM 2015: Drawbridge Cross-Device Connections - Data. https://www.kaggle.com/c/icdm-2015-drawbridge-cross-device-connections/data, 2015.

[6] Measuring Cross-Device, The Methodology. 2016.

[7] Cross-device tracking. Technical report, Federal Trade Commission, 2017.

[8] WebWire - Drawbridge Challenges Scientific Community to Better the Accuracy of Its Cross-Device Consumer Graph. https://www.webwire.com/ViewPressRel.asp?aId=198392, 2017.

[9] https://panopticlick.eff.org/, 2018.

[10] https://easylist.to/, 2018.

[11] https://www.ghostery.com/, 2018.

[12] Automation for Apps. http://appium.io/, 2018.

[13] Customer URL Ticketing System. https://www.trustedsource.org/, 2018.

[14] iSpy? Tailored versus Invasive Ads and Consumers Perceptions of Personalized Advertising. *Electronic Commerce Research and Applications*, 29, 2018.

[15] Pew Research Center - Mobile Fact Sheet. http://www.pewinternet.org/fact-sheet/mobile/, 2018.

[16] Run apps on the Android Emulator. https://developer.android.com/studio/run/emulator/, 2018.

[17] The expert's guide to cross-device conversion & attribution. https://www.tapad.com/uses/the-experts-guide-to-cross-device-conversion-attribution, 2018.

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

[19] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens, and Bart Preneel. Fpdetective: Dusting the web for fingerprinters. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13.

[20] AdBrain. Demystifying cross-device. essential reading for product management,business development and business technology leaders. https://www.iabuk.com/sites/default/files/white-paper-docs/Adbrain-Demystifying-Cross-Device.pdf, 2016.

[21] Adelphic. How cross-device identity matching works. https://adelphic.com/how-cross-device-identity-matching-works-part-1/, 2016.

[22] Elizabeth Aguirre, Dominik Mahr, Dhruv Grewal, Ko de Ruyter, and Martin Wetzels. Unraveling the personalization paradox: The effect of information collection and trust-building strategies on online advertisement effectiveness.

[23] Thakur Raj Anand and Oleksii Renov. Machine learning approach to identify users across their digital devices. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*.

[24] Daniel Arp, Erwin Quiring, Christian Wressnegger, and Konrad Rieck. Privacy threats through ultrasonic side channels on mobile devices. In *(EuroS&P), 2017 IEEE European Symposium on*.

[25] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. Tracing information flows between ad exchanges using retargeted ads. In *25th USENIX Security Symposium*, USENIX Security 16.

[26] Alexander Bleier and Maik Eisenbeiss. Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science*.

[27] Sophie C Boerman, Sanne Kruikemeier, and Frederik J Zuiderveen Borgesius. Online behavioral advertising: A literature review and research agenda. *Journal of Advertising*.

[28] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. A semantic approach to contextual advertising. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07.

[29] Justin Brookman, Phoebe Rouge, Aaron Alva, and Christina Yeung. Cross-device tracking: Measurement and disclosures. *Proceedings on Privacy Enhancing Technologies*.

[30] Xuezhi Cao, Weiyue Huang, and Yong Yu. Recovering cross-device connections via mining ip footprints with ensemble learning. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015.

[31] Yinzhi Cao, Song Li, and Erik Wijmans. (cross-)browser fingerprinting via os and hardware level features. In *Proceedings of Network & Distributed System Security Symposium (NDSS)*. Internet Society, 2017.

[32] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. I always feel like somebody's watching me: Measuring online behavioural advertising. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '15.

[33] Castelluccia C, Kaafar M,Tran M-D. Betrayed by Your Ads! Reconstructing User Profiles from Targeted Ads. In *Privacy Enhancing Technologies*, 2012.

[34] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*.

[35] Hsu-Tang Pu Changfeng C. Wang. Uniquely identifying a network-connected entity.

[36] Kwang Yeun Chun, Ji Hee Song, Candice R. Hollenbeck, and Jong-Ho Lee. Are contextual advertisements effective? *International Journal of Advertising*, 33(2):351–371, 2014.

[37] Criteo. The State of Cross-Device Commerce. `https://www.criteo.com/wp-content/uploads/2017/07/Report-criteo-state-of-cross-device-commerce-2016-h2-SEA.pdf`, 2016.

[38] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015.

[39] Claire Dolin, Ben Weinshel, Shawn Shan, Chang Min Hahn, Euirim Choi, Michelle L Mazurek, and Blase Ur. Unpacking perceptions of data-driven inferences underlying online targeting and personalization. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 493. ACM, 2018.

[40] Drawbridge. Drawbridge Cross-Device Connected Consumer Graph Is 97.3% Accurate. `https://drawbridge.com/news/p/`

drawbridge-cross-device-connected-consumer-graph-is-973-accurate, 2015.

[41] Peter Eckersley. How unique is your web browser? In *Proceedings of the 10th International Conference on Privacy Enhancing Technologies*, PETS'10.

[42] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16.

[43] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L119:1–88, May 2016.

[44] Ayman Farahat and Michael C. Bailey. How effective is targeted advertising? In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12.

[45] Keith Funkhouser, Matthew Malloy, Enis Ceyhun Alp, Phillip Poon, and Paul Barford. Device graphing by example. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[46] Michael C. Grace, Wu Zhou, Xuxian Jiang, and Ahmad-Reza Sadeghi. Unsafe exposure analysis of mobile in-app advertisements. In *Proceedings of the Fifth ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WISEC '12.

[47] Saikat Guha, Bin Cheng, and Paul Francis. Privad: Practical privacy in online advertising. In *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation*, NSDI'11.

[48] Amruta Joshi, Abraham Bagherjeiran, and Adwait Ratnaparkhi. User demographic and behavioral targeting for content match advertising. In *Proceedings of the Fifth International Workshop on Data Mining and Audience Intelligence for Advertising (ADKDD 2011)*, pages 53–60, 2011.

[49] Peter Kafka and Rani Molla. Recode - 2017 was the year digital ad spending finally beat TV. https://www.recode.net/2017/12/4/16733460/2017-digital-ad-spend-advertising-beat-tv, 2017.

[50] Girma Kejela and Chunming Rong. Cross-device consumer identification. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*.

[51] Michael Sungjun Kim, Jiwei Liu, Xiaozhou Wang, and Wei Yang. Connecting devices to cookies via filtering, feature engineering, and boosting. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*.

[52] Aleksandra Korolova and Vinod Sharma. Cross-app tracking via nearby bluetooth low energy devices. In *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy*, CODASPY '18.

[53] Mark Landry, Sudalai Rajkumar S, and Robert Chong. Multi-layer classification: Icdm 2015 drawbridge cross-device connections competition. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*.

[54] Mathias Lécuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. Xray: Enhancing the web's transparency with differential correlation. In *USENIX Security Symposium*, pages 49–64, 2014.

[55] Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15.

[56] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *25th USENIX Security Symposium (USENIX Security 16)*, 2016.

[57] Randall A. Lewis, Justin M. Rao, and David H. Reiley. Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising. WWW '11.

[58] Bin Liu, Anmol Sheth, Udi Weinsberg, Jaideep Chandrashekar, and Ramesh Govindan. Adreveal: Improving transparency into online targeted advertising. In *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*, HotNets-XII.

[59] Lotame. Cross-Device ID Graph Accuracy: Methodology. `https://www.lotame.com/cross-device-id-graph-accuracy-methodology/`, 2016.

[60] Lotame. Cross-device.bridging the gap between screens. `https://www.lotame.com/products/cross-device/`, 2018.

[61] Vasilios Mavroudis, Shuang Hao, Yanick Fratantonio, Federico Maggi, Christopher Kruegel, and Giovanni Vigna. On the privacy and security of the ultrasound ecosystem. *Proceedings on Privacy Enhancing Technologies*.

[62] Jonathan R. Mayer and John C. Mitchell. Third-party web tracking: Policy and technology. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP '12.

[63] Wei Meng, Ren Ding, Simon P Chung, Steven Han, and Wenke Lee. The price of free: Privacy leakage in personalized mobile in-apps ads. In *NDSS*, 2016.

[64] Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[65] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. Privaricator: Deceiving fingerprinters with little white lies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15.

[66] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proceedings of the 2013 IEEE Symposium on Security and Privacy*, SP '13.

[67] Lukasz Olejnik, Tran Minh-Dung, and Claude Castelluccia. Selling off privacy at auction. In *Network and Distributed System Security Symposium (NDSS)*, 2014.

[68] Andriy Panchenko, Fabian Lanze, Jan Pennekamp, Thomas Engel, Andreas Zinnen, Martin Henze, and Klaus Wehrle. Website fingerprinting at internet scale. In *NDSS*, 2016.

[69] Elias P. Papadopoulos, Michalis Diamantaris, Panagiotis Papadopoulos, Thanasis Petsas, Sotiris Ioannidis, and Evangelos P. Markatos. The long-standing privacy debate: Mobile websites vs mobile apps. WWW '17.

[70] Panagiotis Papadopoulos, Pablo Rodriguez Rodriguez, Nicolas Kourtellis, and Nikolaos Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay to reach you? In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17.

[71] Javier Parra-Arnau, Jagdish Prasad Achara, and Claude Castelluccia. Myadchoices: Bringing transparency and control to online advertising. *ACM Trans. Web*.

[72] Patrick Holmes. Mobile and Desktop Advertising Strategies Based on User Intent. `https://instapage.com/blog/adwords-search-device-user-intent`, 2018.

[73] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. Apps, trackers, privacy and regulators: A global study of the mobile tracking ecosystem. NDSS 18.

[74] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, NSDI'12.

[75] Lars Ropeid Selsaas, Bikash Agrawal, Chumming Rong, and Thomasz Wiktorski. Affm: Auto feature engineering in field-aware factorization machines for predictive analytics. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*.

[76] E. Terkki, A. Rao, and S. Tarkoma. Spying on android users through targeted ads. In *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*.

[77] Vincent Toubiana, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas. Adnostic: Privacy Preserving Targeted Advertising. NDSS '10.

[78] Catherine E Tucker. Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research*, 51(5):546–562, 2014.

[79] Jeremy Walthers. Learning to rank for cross-device identification. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*.

[80] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09.

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

[82] Sebastian Zimmeck, Jie S. Li, Hyungtae Kim, Steven M. Bellovin, and Tony Jebara. A privacy analysis of cross-device tracking. In *26th USENIX Security Symposium*, USENIX Security 17, pages 1391–1408. USENIX Association, 2017.

That is not dead which can eternal lie,
And with strange aeons even death may die...

The Nameless City
-H.P.Lovecraft