Automated Measurements of Cross-Device Tracking

Konstantinos Solomos,[1] Panagiotis Ilia,[1] Sotiris Ioannidis,[1] Nicolas Kourtellis[2]

[1] FORTH, Greece
{solomos, pilia, sotiris}@ics.forth.gr
[2] Telefonica Research, Spain
nicolas.kourtellis@telefonica.com

Abstract. Although digital advertising fuels much of today's free Web, it typically do 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 effectiveness of ads, advertisers have introduced new paradigms such as cross-device tracking (CDT), to monitor users' browsing on multiple screens, and deliver (re)targeted ads in the appropriate screen. Unfortunately, this practice comes with even more privacy concerns for the end-user. In this work, we design a methodology for triggering CDT by emulating realistic browsing activity of end-users, and then detecting and measuring it by leveraging advanced machine learning tools.

1 Cross-Device Tracking: An Imminent Privacy Problem

Online advertising can be easily tailored to the audience, to become personalized to each particular user according to her needs and interests. Until recently, ad-companies would typically target each user with ads according to the behavior on a specific device. However, since users own multiple devices, advertisers started moving towards more advanced targeting practices that are designed to track users across their devices, and target users regardless of the device used.

According to a recent FTC Report [1], Cross-Device Tracking (CDT) can be deterministic, where first-party login services (e.g., Facebook, Gmail) that can track users across devices are being used, or probabilistic, where there are no shared identifiers between devices, and third parties try to identify which devices belong to the same user by considering network access data and patterns in browsing history and behavior etc. Ad-companies that engage in CDT typically use a mixture of both techniques, but in either case, the implications for user privacy are severe: they are capable of tracking users across all their digital space, and use such information in a non-transparent fashion.

It is inherently difficult to detect and measure probabilistic cross-device tracking in a systematic way, as it is heavily based on user activity. Therefore, recent privacy regulations (e.g., EU's GDPR and ePrivacy) will not be easy to enforce in such cases. The main problem in measuring CDT lies in distinguishing which ads are presented to the user because of her behavior on a device (targeting or retargeting), and which ads are because of her activity on a different device.

A few recent works investigated CDT based on technologies such as ultrasound and Bluetooth, and measured the prevalence of these approaches [2,3]. A work by Brookman et al. [4], one of the few that investigate CDT on the web, provides some initial insights about the prevalence of trackers. It examines 100 popular websites in order to determine which of them disclose data to trackers and identifies which websites contain trackers known to employ CDT techniques. Zimmeck et al. [5] designed an algorithm that estimates similarities and correlates the devices into pairs, based on IP addresses and browsing history. That approach shows that users' network information and browsing history can be used for pairing user devices, and thus potentially for CDT.

Our work builds on these early studies on CDT and also on past studies on the detection of web tracking during targeted ads. We propose a first of its kind methodology for the systematic investigation of probabilistic CDT, by leveraging artificially-created behavioral profiles, and measuring the factors affecting CDT in various experimental setups. The contributions of this work are as follows:

- A methodology for detecting CDT based on triggering behavioral cross-device targeted ads on one user device, according to a specific emulated browsing behavior, and then detecting these ads when delivered on a different device.
- An investigation of the factors that affect the performance of CDT under different experimental configurations. We establish artificially-created behavioral profiles with specific web behaviors, and measure the existence of CDT.

2 Methodology to measure CDT

The main goal of this work is to design a concrete methodology for measuring cross-device tracking activity, as well as to identify the dominant factors that affect the performance of CDT. The methodology emulates realistic browsing activity of end-users with specific web interests across different devices, and collects and analyzes all ads delivered to these devices, due to static advertising or targeted, behavioral or retargeted advertising. Finally, it compares these ads with baseline browsing activity to establish if cross-device tracking is present or not, at what level, its lifetime, and for which types of user interests.

The design of this methodology focuses on the following design objectives:

- Ability to detect probabilistic CDT in a systematic and repeatable fashion.
- Scalability, for fast deployment of multiple parallel device instances, for increased data collection.
- Support the investigation of cross-device tracking in both directions, i.e., mobile → desktop, and desktop → mobile.
- Employ advanced machine learning analysis to compute probability of cross-device tracking in a given experimental setup.
- Support short and long-term experiments, for data collection in ad-hoc fashion or historically through time, respectively.

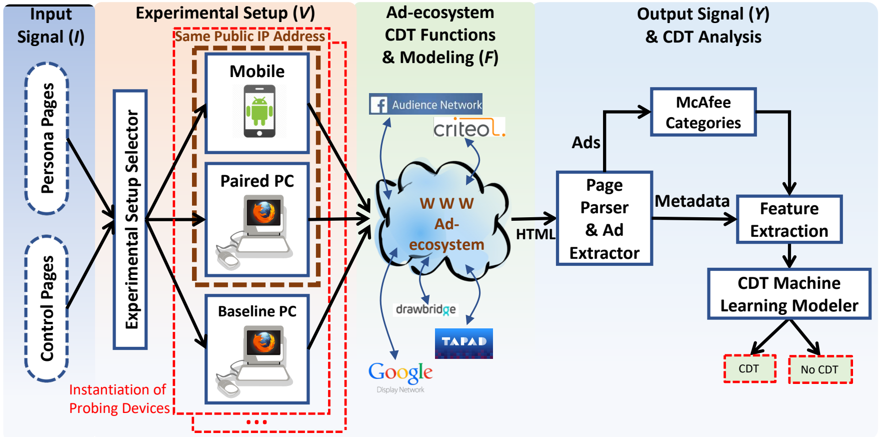

Fig. 1. High level representation of methodology design principles and units.

2.1 Design Principle

In general, we consider the cross-device tracking performed by the ad-ecosystem as a complex process, with multiple parties involved, and not easy to dissect and understand. To infer its internal mechanics we rely on probing this ecosystem with consistent and repeatable inputs ($\mathcal{I}$), under specific experimental settings ($\mathcal{V}$), allow the ecosystem to process and use this input via transformations and modeling ($\mathcal{F}$), and produce measurable outputs on the receiving end ($\mathcal{Y}$):

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

Following this design principle, our methodology allows researchers to push realistic input signals to the ad-ecosystem via website visits, and measure the ad-ecosystem's output through the delivered ads, to demonstrate if $\mathcal{F}$ has allowed or not the ad-ecosystem to perform probabilistic CDT. Based on this design principle an overview of our methodology is illustrated in Figure 1.

2.2 Methodology Challenges & Considerations

Devices and IPs. The approach we follow is based on triggering and identifying cross-device targeted ads, specifically ads that appear on one of the user's devices but have been triggered by the user's activity on a different device. Our methodology requires a minimum of three different devices (as seen in Figure 1): one mobile device and two desktop computers, as well as two different IPs. We assume that two of these devices (i.e., the mobile and one desktop) belong to the same user and are connected to the same network. 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 paired PC.

Emulating user behavior with personas: Training Phase. To trigger CDT, we first need to input to the ad-ecosystem some network 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, the method visits specific websites to emulate a user's behavior according to some predefined *personas*, similarly to Carrascosa et al. [6]. We leverage this approach for emulating browsing behavior according to specific user interests (e.g., travel and vacations, sports, shopping, etc.), and to create multiple personas of different granularities. For each of the personas, the methodology can identify a set of websites that have active ad campaigns (*training pages*), which the given persona visits and interacts with during the *training* phase.

Control pages: Testing Phase. To reduce any bias from possible behavioral ads delivered to specific type of websites, the desktops collect ads by visiting *control pages*, i.e., neutral websites (weather, news) that typically serve ads not related to their content. During the testing phase, each device visits the control pages, and the method extracts, analyzes and categorizes the collected ads, in order to identify those ads that have been served to the user's desktop computer because of the browsing behavior on the mobile device.

CDT Detection: Comparing Signals. In order to detect CDT, 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 desktop. For example, methods that perform similarity computation between the two signals in a given dimensionality (e.g., Jaccard, Cosine) can be of use. However, since the ad distribution techniques used by the ad-ecosystem are still unknown, we can employ advanced methods, such as machine learning techniques, for the classification of the signals as similar enough to match, or not, based on specific features from the experimental setup ($\mathcal{V}$), and the input/output features. In this methodology we opt for typical methods of machine learning, to compute the likelihood of the two signals being the product of CDT.

3 Measuring CDT in the Wild

This section describes the operational settings of our methodology during the experiments conducted for measuring CDT and its effect on the ad-ecosystem.

3.1 Experimental Setup

Personas & Training Pages. A critical part of the methodology is the design and automatic building of realistic user personas. Each persona has a unique collection of visiting websites that form the set of *persona pages*. Since we do not

Table 1. Behavioral personas generated 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.

know in advance which e-commerce sites are conducting cross-device campaigns, our personas must cover a wide area of interests. For this reason, we use the persona categorization of Carrascosa et al. [6] for the top 50 personas, and resolve the taxonomy list[1] to obtain the related keywords. We group by the taxonomy keywords based on their content and then we form sets of labels describing the personas. For capturing active ad campaigns we use Google Search, as it reveals ad campaigns associated with products currently being advertised. That is, if a user searches for specific keywords (e.g., "men watches"), Google Search will provide a set of results, including a list of sponsored links from e-commerce sites and services conducting campaigns for the terms searched. This procedure is repeated until at least five, and a maximum of ten, unique domains per persona are collected from the Google Search results.

In general, our method is able to generate a large number of different personas, corresponding to various interests and online behaviors: from generic to specific taxonomy categories. As the effectiveness of a persona depends on the active ad campaigns at time, in our experiments we deploy only the 10 personas shown in Table 1.

Experimental Settings. Each experiment is executed multiple times (or runs), through parallel instantiations of the user devices within the implemented methodology. Each experimental run is executed following a timeline of phases as illustrated in Figure 2. This timeline contains N sessions with three primary stages in each: 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 into 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 device during training with a persona, but also performs a subsequent test browsing with the control pages to establish the state of ads after the training. Finally, the *After*

[1] `https://www.google.com/basepages/producttype/taxonomy.en-US.txt`

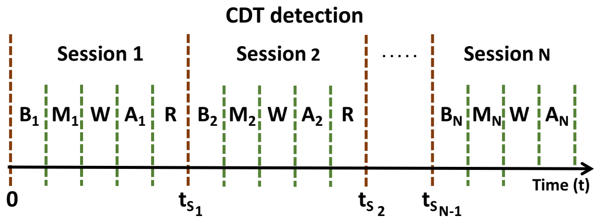

Fig. 2. Timeline of phases for CDT measurement. M_i: mobile training; $B_i(A_i)$: testing time before (after) mobile training; W(R): wait (rest) time.

(A_i) stage is when the two desktop device perform the final test browsing to establish the state of ads after the mobile training.

After in-depth experimentation, we found that training time t_{train}=15 minutes and testing time t_{test}=20 minutes are enough for creating, collecting and processing a satisfying number of data without introducing noise to the web traffic, while keeping clear each device's signal to the ad-ecosystem. There is also a waiting time (t_{wait}=10 minutes) and resting time (t_{rest}=5 minutes) between the stages of each session, to allow 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.

Machine Learning Algorithms and Performance Metrics. Our analysis is based on three classification algorithms with different dependence on the data distribution. An easily applied classifier that can be used for performance comparison with the other models in a baseline fashion is Gaussian Naive Bayes. Logistic Regression is a well-behaved classification algorithm that can be trained as long as the classes are linearly separable. At last, Random Forest is a widely used 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.

A critical point when considering the performance evaluation of these algorithms is the selection of the appropriate metrics, since we want to report the most accurate estimation for the number of predicted paired devices, while at the same time measure the absolute number of misclassified samples overall. For this reason, metrics like Precision, Recall and F_1-score are typically used, since they can quantify this type of information. There is also one more metric used for measuring the dependence of the True Positive Rate (TPR) with the False Positive Rate (FPR). If we plot the curve of those two rates for different operational scenarios, we get the Receiver Operating Characteristic curve (ROC). If a single numeric score based on the ROC curve is needed, then the Area Under the Curve (AUC) is used.

3.2 Detecting CDT

This experimental setup emulates a realistic user behavior browsing frequently about some specific topics, but in short-lived sessions in her devices. Given that

Table 2. Performance evaluation for Random Forest. Left value in each column is the score for Class 0 (C0=*not paired desktop*); right value for Class 1 (C1=*paired desktop*).

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

most users do not frequently delete their local browsing state, this setup assumes that the user's browser keeps all state, i.e., cookies, cache files, browsing history, etc. This assumption enables trackers to identify users easier across their devices, as they have historical information about these users. All 10 personas of Table 1 are used, while the data collection for each Persona lasts ∼4 days.

The classification results for the Random Forest algorithm, as reported in Table 2, had the best performance compared to the other two algorithms. We use AUC score as the main metric score, since the ad-industry seems to prefer higher Precision scores over Recall, as the False Positives have greater impact on the effect of ad-campaigns. As shown in Table 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.

In order to retrieve the variables that affect the discovery and measurement of cross-device tracking, we applied the feature importance method on the dataset of each persona, and selected the top-10 highest scoring features. 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 some cases, there were also landing pages that had high scoring, but this was not consistent across all personas.

These results indicate that for high scoring personas, we successfully captured the active CDT campaigns, but for the personas with lower scores, there may not be active cross-device tracking campaigns for the period of the experiments. Finally they also 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 the delivered ads.

4 Outcome and future directions

Undoubtedly, CDT has an 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. To this direction we proposed a concrete and scalable methodology that allows experimenting with different CDT scenarios. We plan to extend this work by designing and conducting various different experiments that will shed light and help understand the mechanics behind CDT as applied by the complex ad-ecosystem. Furthermore, the extensibility of the platform enables using new methods invented in the future for better input signal generation (e.g., for persona building), device emulation with more realistic browsing behavior, more precise webpage parsing for ads extraction, new machine learning algorithms etc. The proposed methodology will also enable the community of privacy researchers and advocates to study CDT in a systematic way, and to quantify its intensity and impact to users with different, and potentially sensitive or legally protected web interests and online behaviors.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation programme under Grand Agreement No 700378 (project CIPSEC) and the Marie Sklodowska-Curie Grand Agreement No 690972 (project PROTASIS). This paper reflects only the authors' view and the Agency is not responsible for any use that may be made of the information it contains.

.

References

1. FTC: Cross-device tracking. Technical report (2017)
2. Mavroudis, V., Hao, S., Fratantonio, Y., Maggi, F., Kruegel, C., Vigna, G.: On the privacy and security of the ultrasound ecosystem. Proceedings on Privacy Enhancing Technologies **2017**(2) (2017) 95–112
3. Arp, D., Quiring, E., Wressnegger, C., Rieck, K.: Privacy threats through ultrasonic side channels on mobile devices. In: Security and Privacy (EuroS&P), 2017 IEEE European Symposium on, IEEE (2017) 35–47
4. Brookman, J., Rouge, P., Alva, A., Yeung, C.: Cross-device tracking: Measurement and disclosures. Proceedings on Privacy Enhancing Technologies **2017**(2) (2017) 134–149
5. Zimmeck, S., Li, J.S., Kim, H., Bellovin, S.M., Jebara, T.: A privacy analysis of cross-device tracking. In: 26th USENIX Security Symposium. USENIX Security 17, Vancouver, BC, USENIX Association (2017) 1391–1408
6. Carrascosa, J.M., Mikians, J., Cuevas, R., Erramilli, V., Laoutaris, N.: 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, New York, NY, USA, ACM (2015) 13:1–13:13