Selling Off Privacy at Auction

Lukasz Olejnik
INRIA
lukasz.olejnik@inria.fr

Minh-Dung Tran
INRIA
minh-dung.tran@inria.fr

Claude Castelluccia
INRIA
claude.castelluccia@inria.fr

***Abstract*—Real-Time Bidding (RTB) and Cookie Matching (CM) are transforming the advertising landscape to an extremely dynamic market and make targeted advertising considerably permissive. The emergence of these technologies allows companies to exchange user data as a product and therefore raises important concerns from privacy perspectives. In this paper, we perform a privacy analysis of CM and RTB and quantify the leakage of users' browsing histories due to these mechanisms. We study this problem on a corpus of users' Web histories, and show that using these technologies, certain companies can significantly improve their tracking and profiling capabilities. We detect 41 companies serving ads via RTB and over 125 using Cookie Matching. We show that 91% of users in our dataset were affected by CM and in certain cases, 27% of users' histories could be leaked to 3rd-party companies through RTB.**

We expose a design characteristic of RTB systems to observe the prices which advertisers pay for serving ads to Web users. We leverage this feature and provide important insights into these prices by analyzing different user profiles and visiting contexts. Our study shows the variation of prices according to context information including visiting site, time and user's physical location. We experimentally confirm that users with known history are evaluated higher than new comers, that some user profiles are more valuable than others, and that users' intents, such as looking for a commercial product, are sold at higher prices than users' browsing histories. In addition, we show that there is a huge gap between users' perception of the value of their personal information and its actual value on the market. A recent study by Carrascal et al. showed that, on average, users evaluate the price of the disclosure of their presence on a Web site to EUR 7. We show that user's browsing history elements are routinely being sold off for less than $0.0005.

I. Introduction

Online advertising is prevalent on the Web and brings substantial revenues to a majority of Internet companies. Consequently, increasingly sophisticated methods, often based on complex analysis of users' data, have been developed to improve the efficiency of advertising.

Real Time Bidding (RTB) [19] is a novel paradigm of serving ads with the aim of bringing more liquidity to the online advertising market. When a user visits a Web site which displays advertisements (ads) through RTB, the ad request is sent to an Ad Exchange which subsequently broadcasts it along with user data to *ad buyers* and holds an auction. These buyers bid in this auction and the winning party is allowed to serve ads to the user. The underlying technology to exchange users' identification data between Ad Exchanges and buyers is *Cookie Matching*, which allows two different domains to match their cookies of the same user.

Although RTB and Cookie Matching are acclaimed by the advertising industry, their privacy implications are not adequately understood. Cookie matching enables the possibility of linking the profiles of a single user in databases of two independent companies and is an integral part of RTB. In RTB, Ad Exchanges leverage Cookie Matching to broadcast user data to ad buyers. In other words, users' data become a product that is auctioned in real time in the online advertising market.

RTB-based spending is growing rapidly and is expected to account for more than 25% of the total display advertising sales in the US by 2015, up from 10% in 2011. By 2015, the majority of indirect display ad sales revenue will be traded using RTB in the United States and the most developed European markets [26]. RTB and Cookie Matching become increasingly rampant in the online advertising industry, yet to the best of our knowledge, there have been little academic studies of their privacy implications. In this paper, we conduct an empirical study of these technologies and analyse how they impact users' privacy. We believe that it is important for users, researchers and privacy advocates to understand this privacy implication in very details.

While estimating value of user's private information is an interesting problem [4], [5], evaluating it is subtle and not obvious. Several recent research studies established results from the users' perspective [7]. Users, however, often do not have a developed sense of privacy. We approach this problem from the advertisers' perspective based on a market principle: *users' private data are worth as much as someone is willing to pay for them*. By leveraging a design feature of RTB systems, we are able to observe prices that advertisers pay for an ad impression after winning an auction. We utilize these prices to conduct a detailed analysis of the value of users' private data, with a focus on users' Web browsing history.

In summary, our main contributions in this paper include:

- We quantify the impact of Cookie Matching (CM) and Real-Time Bidding (RTB) on users' privacy. We show that CM happens very frequently and is performed by a large number of companies; some of them execute Cookie Matching in a significant proportion of the studied users' profiles (up to 91% of the 100

NDSS '14, 23-26 February 2014, San Diego, CA, USA
 ISBN 1-891562-35-5
http://dx.doi.org/doi-info-to-be-provided-later

profiles we studied in our experiments). Our analysis of RTB shows that Ad Exchanges (e.g. DoubleClick) broadcast user-visited sites to a considerable number of bidders in real time; some of the bidders can learn up to 27% of users' histories through this mechanism.

- We provide an analysis of the value of users' private data from the advertisers' perspective based on prices they paid for serving ads to users. We analyze how such factors as the visiting site, time, user's physical location and user's profile affect prices actually paid by advertisers. Interestingly, we discovered that prices are highest in the early morning. Prices in the US (average $0.69 CPM, an equivalent of $0.00069) are observably higher than those in the cases of France ($0.36 CPM) and Japan ($0.24 CPM). We confirm the fact that when a user's Web history is previously known to advertisers, they are willing to pay a higher price than in the case of new users. We also show that users' intents, such as browsing a commercial product, are higher valuated than their general histories, i.e. browsing sites not related to specific products. Finally, we highlight a huge gap between users' perception of the value of their personal information and its actual value on the market. In fact, a recent study by Carrascal et al. [7] indicated that on average, users evaluate the price of the disclosure of their presence on a Web site to EUR 7. We show that this piece of data is actually being sold off for strikingly lower prices: less than $0.0005.

The rest of the paper is organized as follows. We provide background information in section II. We then present techniques used to detect Cookie Matching and Real-Time Bidding in section III. We give detailed analysis on Cookie Matching and Real-Time Bidding in section IV, then analyze the winning prices in RTB auctions in section V. We discuss privacy and related problems in section VI. We present related work in section VII and conclude in section VIII.

II. Background information

A. Cookie Matching

Cookie Matching (CM), an integral part of Real-Time Bidding, is a mechanism allowing two separate parties to synchronize their users' cookies [24]. For example, an Ad Exchange and a Bidder (ad buyer) normally attribute their own distinct cookies to the same user. After an execution of Cookie Matching protocol, one or both of them will have these cookies mapped to each other. Some Ad Exchanges, notably DoubleClick, create and use a unique user id (e.g. one-way hash of the cookie) instead of a cookie, with the aim to protect the actual cookie content from being revealed to the Cookie Matching partners. Nevertheless, we detected that many others are sending clear-text cookies for matching.

Figure 1 shows the main phase of Cookie Matching. Ad Exchange typically sends a script or a redirect instruction in order to instruct the user's browser to load a URL provided by the Bidder with the Ad Exchange's user's cookie/id in the parameter. The Bidder obtains the Ad Exchange's cookie/id upon receiving this request and matches this cookie/id with its own cookie. In some cases, this process can happen in the reverse direction, which results in the Cookie Matching on the Ad Exchange's side. Cookie matching is also known under different names, such as cookie syncing, pixel matching, etc. In this paper, we use "Cookie Matching" for all such actions of cookie synchronization between two separate entities.

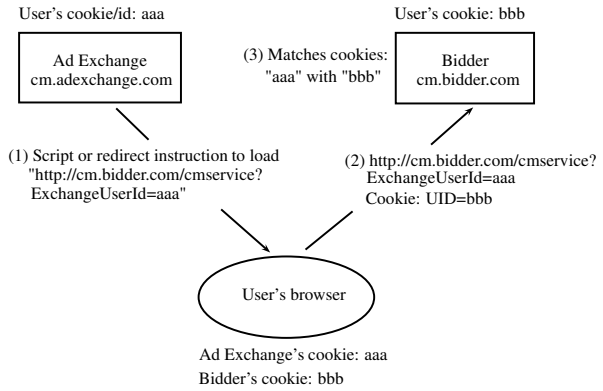


Fig. 1: Cookie matching protocol

B. Real-Time Bidding

Real-Time Bidding (RTB) [19] allows advertisers to buy online advertisement spaces at real-time through Ad Exchanges. Here we discuss the mechanism of DoubleClick's Ad Exchange [12], which is likely the most representative. Other Ad Exchanges, for example Pulse Point [35], employ similar approaches. The OpenRTB initiative [25], which aims at standardizing RTB, provides a similar description.

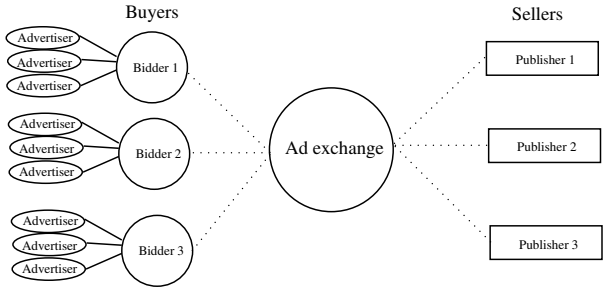


Fig. 2: Ad Exchange model

The four main entities taking part in Real-Time Bidding include: 1) *Publishers* (e.g. nytimes.com), possessing the Web sites which display ads, 2) *Ad Exchanges* (e.g. DoubleClick), which enable ad transactions between ad sellers (the *publishers*) and ad buyers (the *bidders*) based on auctions held in real time, 3) *Bidders* (e.g. Criteo[1]), which are big advertising agencies representing small and medium advertisers to bid in RTB auctions in order to win ad spaces, and finally 4) *Advertisers* (e.g. hotels.com), which want to advertise and sell their products or services. Each time an ad is displayed in a Web site visited by a user, we call this event an *ad impression*. The RTB mechanism works as follows: When a user visits a publisher's Web site belonging to an Ad Exchange's advertising network, a HTTP request is sent to

[1] http://www.criteo.com

the Ad Exchange. The Ad Exchange subsequently sends *bid requests* for this ad impression to its bidders. Bidders then analyze the impression and submit their *bid responses*, which include prices they are willing to pay and their ad snippets. The bids are submitted to an online auction, and the Ad Exchange serves the winner's ad snippet on the user's visited Web site. After a successful transaction, the Ad Exchange charges the winning bidder, and pays the publisher after subtracting a commission. The total process generally happens in less than $100\,ms$.

Bid requests sent from Ad Exchanges to bidders typically contain information such as the Ad Exchanges' user's cookie (or user's id) and the user's visiting context including the following information: the URL of the Web site being visited, the categories of the site, the first three bytes of the user's IP address[2], various information concerning the user's browser and others [13], [14]. Upon receiving a bid request, the bidder finds its user cookie through the Ad Exchange's cookie thanks to Cookie Matching, provided that this protocol has been executed previously. It then determines the bid price based on the user's profile it possesses and the user's context provided by the Ad Exchange. Bidders can also bid on new users about whom they do not possess any prior information. When a bidder wins the auction, it has the right not only to serve ads, but also to initiate a Cookie Matching with the Ad Exchange.

An online Ad Exchange works similarly to a stock exchange [46], only trading in audiences for online ads. This mechanism helps publishers to sell their ads at the most competitive price, while allowing bidders to flexibly adjust their buying strategy in real time.

C. The Economics of Real-Time Bidding

The payment model used in Real-Time Bidding is Cost-per-mille impression (CPM) [21], which means every transaction through Ad Exchange is on a pay-per-impression basis. However, some advertisers might prefer the Cost-per-click (CPC) model [20], as its performance is more effectively measurable than CPM. As a result, a hybrid model exists, in which real-time bidders (e.g. Criteo) buy ad impressions from Ad Exchanges and sell ad clicks to advertisers. In this model, the bidders are expected to bid high enough in Ad Exchange in order to win the auction, while ensuring an adequate click probability to gain a margin benefit. Click probability depends largely on how the ad content matches user profiles and/or visiting contexts.

In this work, we aim to analyze how bidders evaluate users' personal data on behalf of advertisers. We therefore focus on analyzing the strategy from the advertiser's perspective. Advertiser's purposes normally include: 1) inviting users to their Web sites for buying a product or using a service, and/or 2) improving brand awareness. In both cases, the common goal is to reach potential customers. As most of Ad Exchanges encourage truthful bidding, for example by the use of Vickrey auctions [43], the best strategy for advertisers is expected to be bidding in accordance with the true value they can expect to get from the user.

[2]Note that some companies, such as Pulse Point, actually send full IP addresses [36].

III. Cookie Matching and RTB Detection

In this section, we describe the discovery techniques that we employed. First, we introduce the request hierarchy detection technique, which serves as a basis for the others. Second, we present our technique to detect Cookie Matching. Then we describe the Real-Time Bidding detection technique, which is based on the discovery of winning prices.

A. Request hierarchy detection

We describe our technique to detect all causal relations between HTTP requests. The requests are often originating from Web sites' HTML tags including `<script>`, `<img>` or `<iframe>`. The responses to these requests might also contain HTML elements or JavaScript code that subsequently initialize other requests, and so on. Detecting such causal relations between requests is important to observe Cookie Matching and Real-Time Bidding events.

Assuming two HTTP requests A and B, A happening before B, our approach is as follows: we observe the *HTTP Referer* field in the request header of B (B's Referer), and *Location* field in the response header of A (A's Location). If A's Location contains B's URL, this means the browser redirects the request from A to B. Meanwhile, B's Referer containing A's URL means B is loaded from the content of A. Nevertheless, in the case of requests being dynamically initiated as a result of JavaScript scripts, the Referer field might not be a good indicator, as it points to the visited Web site rather than the source of the script. Therefore, we also scan all the JavaScript files we encounter during the loading of the site. If a request's URL is detected in a JavaScript script, we conclude that the script creates this request. However, this approach fails when JavaScript code builds URLs dynamically by concatenating dynamic parameters into a domain. We therefore also search JavaScript scripts for domains.

B. Cookie matching detection

In Cookie Matching, one domain synchronizes its cookie with another domain by including it in the request sent to the latter. For example, domain A returns a script to the browser which will invoke a request to domain B such as: http://B_URL?ExternalUserId=[A's cookie] (see section II-A). Therefore, in order to detect Cookie Matching, we detect all the causal relationship $A \rightarrow B$, then scan all cookies from A and all parameters sent to B. We only take into account values that are sufficiently long, i.e. whose length exceeds 10 characters, as shorter strings are usually temporary values, unrelated to our research. If a match is detected, we consider it to be Cookie Matching. We manually checked a considerable number of values to confirm that they are indeed long-term cookies.

This method fails with DoubleClick, as this company uses a unique *user id* instead of the cookie itself. In this case, we leverage the Google's Cookie Matching protocol description [24], which clearly defines specific URL patterns. Examples of these URLs are presented in Table I. In these URLs, `google_nid` is the unique id that Google assigns to its Cookie Matching partner (ad buyer), while `google_gid` is the Google user id corresponding to the Google's user's cookie.

TABLE I: Google's Cookie Matching URLs

Google's Cookie Matching URLs
http://cm.g.doubleclick.net/pixel?google_nid=[...]&google_cm
http://cm.g.doubleclick.net/pixel?google_nid=[...]&google_push=...

Google distributes buyer-specific user ids, which means different buyers see different Google user ids for the same Web user.

C. Real-Time Bidding detection

Bidders are charged for every ad impression won through RTB. The paid prices are usually included in the requests related to ad creatives which are served via Ad Exchanges with the help of a `WINNING_PRICE` macro. We detect RTB by interpreting the values of parameters in HTTP requests and looking for such price pattern.

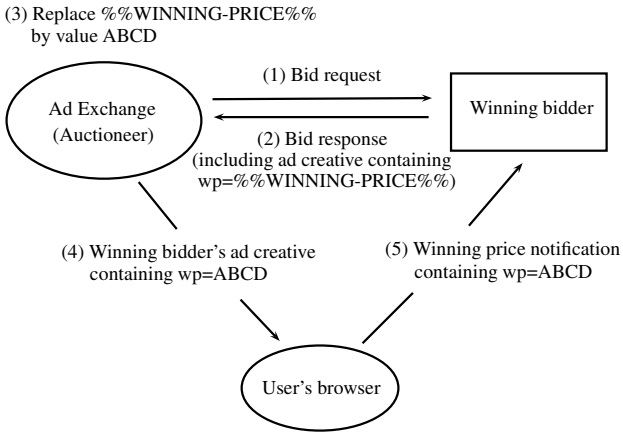


Fig. 3: Winning price notification in Real-Time Bidding

The purpose of the `WINNING_PRICE` macro is to allow the Ad Exchange to notify the winning bidder about the actual price it has to pay for the ad impression[3]. The mechanism of winning price notification is shown on Figure 3. In its bid response sent to the Ad Exchange, each bidder includes its ad creative, i.e. a small HTML or JavaScript code responsible for displaying ads. The ad creative normally contains the winning price macro in a special text form (e.g. *%%WINNING_PRICE%%* in the case of DoubleClick) appended to a URL (which we call *ad_URL*). After the auction, the Ad Exchange replaces the winning price macro in the winner's ad creative with the actual winning price, and serves the creative to the user. Upon reception of this message, the user's browser runs the creative which initializes a HTTP request to the *ad_URL* in order to fetch the actual advertisement. Note that this HTTP request also contains in its parameters the winning price.

In the following, we describe the DoubleClick's *winning price format*. During the experiments we conducted in this work (section IV and V), we detected a considerable number of other companies apparently using the same or similar formats. In DoubleClick Ad Exchange, which belongs to Google, the price is encrypted and subsequently has a fixed length of 28 bytes (Figure 4). It is then encoded in a 38-character-length Web-safe Base64 [22].

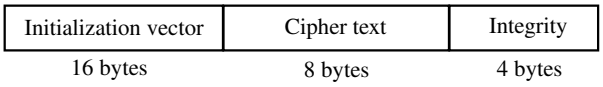


Fig. 4: Google's winning price format

Each bidder shares a different *encryption key* and *integrity key* with the Ad Exchange to allow the decryption and verification of the encrypted price. The *initialization vector* contains a timestamp in the first 8 bytes with the aim to detect any stale response attack [22]. We rely on this timestamp to detect encrypted prices. Specifically, we extract all the suspected values of the URL's parameters, which have a length of 38 characters and contain only valid Web-safe Base64 characters, in each HTTP request initiated by the browser. We decode each of these values and extract the first 8 bytes to investigate whether this is a valid timestamp. According to Google's description, we convert the first 4 bytes to seconds and the last 4 bytes to milliseconds, and then obtain the total milliseconds. We compare this timestamp to the timestamp obtained from the response header of the investigated request. If they do not differ beyond a threshold, we consider the timestamp as valid, and assume the encrypted text is a valid price. We use a 5-minute threshold.

Each price is included in a URL as a value of a specific parameter. For example, the creative could include a URL in the following form: http://bidder_URL?wp=[Winningprice], here `wp` being a URL's parameter whose value is the winning price. We used the winning price detection technique described previously to detect such forms of URLs. Table II provides some examples of the domains and the corresponding parameter names we encountered during our experiments and tests. For example, the price URL for Invite Media has the following form: http://invitemedia.com?cost=[Winningprice] (extra parameters stripped for clarity) – the price parameter is `cost` in this case.

When investigating requested URLs during our experiments in search for such patterns, we surprisingly found a substantial number of winning prices that were not encrypted. We deduce that these values are winning prices because of the following reasons. First, these URLs share identical patterns with URLs containing encrypted prices (same domain name, same list of parameters), but include a clear-text value instead of an encrypted one for the same URL's parameter. Second, the values we obtained were very often in form of floating-point number (e.g. 0.5) or integer in micros format (i.e. 1 is converted to $1,000,000$ micros), which match exactly the price format description of the advertising industry. Moreover, the parameters' names for these values are often contextual and meaningful. Examples include: "*win_price*", "*cost*", "*price*" or even "*rtbwinprice*" as shown in Table II. In total, we detected 41 domain names (e.g. *ad.turn.com*) belonging to advertisers (*Turn* in this case), and corresponding HTTP parameters (*acp* in this case) whose values contained prices.

[3]This price is not necessarily equal to the actual bid price as most of Ad Exchanges use second-price auctions, in which the winner pays the *second highest* bid price incremented by a small pre-defined value.

TABLE II: Clear-text price URL patterns

Domain	Parameter name
invitemedia.com	cost
mathtag.com	price
gwallet.com	win_price
adnxs.com	pp
mythings.com	rtbwinprice

It is understandable that companies use the same URL patterns for winning price notification, regardless of the formatting of prices, while working with different Ad Exchanges. This helps them maintain a unified and simpler information system. The fact that a significant proportion of prices are in clear-text gives us an opportunity to observe how advertisers evaluate the value of each impression (see Section V).

In summary, we use both encrypted and clear-text prices to detect Real-Time Bidding. It should be noted that winning price notification is not obligatory. Rather, it is an option for bidders and depends on the policy of Ad Exchanges. This means there might be some communications related to Real-Time Bidding that we could not detect. This happens if Ad Exchanges choose other schemes to notify the winning prices (e.g. server to server) or real-time bidders do not use the `WINNING_PRICE` macro. Therefore, the number of Real-Time Bidding communication we detected using this scheme can be considered as a *lower bound* of the actual number.

IV. Cookie matching and RTB analysis

A. The RTBAnalyser plugin

We implemented all the aforementioned techniques in a Firefox plugin, *RTBAnalyzer*, which is a modified version of HttpFox [3], an open source Firefox plugin. We implemented a Firefox `nsIObserver` interface to observe all HTTP requests and responses, then applied the previously-described techniques to detect Cookie Matching and Real-Time Bidding. The plugin builds a hierarchy organization of all HTTP requests originating from the sites visited by the user. For each request, it collects the domain name (not the full URL) and identify whether the request is related to Cookie Matching or Real-Time Bidding. In case of Real-Time Bidding, it also collects the related winning prices. These analyzed information are saved into JSON format and sent to our server. It is important to note that each domain name of first-party sites contained in these data reports is replaced with a random value in order to protect privacy of plugin users.

B. Dataset

We distributed the plugin to our colleagues and friends and asked them to install it and browse the Web normally for a number of days. The experiment was performed during the month of June, 2013. The volunteers were mostly researchers and students based in our country of residence, France. Data were automatically sent to our servers every hour or at user request, depending on the chosen installation option. We did not attempt to create any link between the data we obtained and the personal identities of the users. As a result, we do not know who actually participated in the experiment. At the end of the experiment, we selected the top 100 profiles, after removing those that contained less than 70 sites. This dataset is used in sections IV-C and IV-D, and part of section V.

C. Cookie matching privacy analysis

1) Privacy analysis: Companies normally build independent user profiles identified by their own cookies. Cookie Matching facilitates potential cooperation between these systems to exchange their users' data and possibly build larger user profiles. Without matching cookies, it would be difficult to link two profiles of the same user maintained by two separate entities. This results from the fact that trackers are usually able to see only the URLs a user is visiting and no other identifying information, such as e-mail address or user's name. While tracking and data exchange for advertising purposes are increasingly prevalent, technologies like Cookie Matching could potentially enable user tracking to a much larger scale.

2) Methodology: In order to demonstrate and quantify the potential risks described in the previous section, we studied the 100 profiles and identified the most active companies performing CM. Simultaneously we monitored the top trackers of these profiles, and evaluated the extent of potential history discovery by these entities via tracking. Finally, we evaluated to what extent these companies could broaden their tracked users' profiles by making use of CM and sharing their knowledge of profiles.

3) Results: We first counted the cumulated numbers of Cookie Matching events following each site of the profiles in the real user dataset, and then averaged out these values. We show the results for the first 70 sites in the users' histories. Figure 5 displays these average values according to the number of visited sites. It shows that more than 60 Cookie Matching events happen when a user visits 40 sites (red curve) and more than 30 domains are involved (green curve). We can observe that the number of Cookie Matching increases regularly with the number of visited sites. The average cumulated number of cookie matching events after 70 visited sites is more than 100, performed by nearly 60 different domains on average. These results show that Internet users are encountering Cookie Matching at regular intervals.

We observed the frequency of Cookie Matching performed by each pair of companies, and detected that many of them executed this scheme routinely. Table III shows the 20 pairs of domains that performed Cookie Matching the most. We noticed that Facebook (facebook.com) and AppNexus (adnxs.com) matched their cookies in 91% of profiles. The numbers are respectively 87%, 86% and 85% for the following pairs: Turn (turn.com) - Admeld (admeld.com), DoubleClick (doubleclick.net) - Rfihub (rfihub.com) and DoubleClick (doubleclick.net) - AppNexus (adnxs.com).

We investigated the *top* 25 *trackers*, i.e. the domains that tracked the largest parts of the studied users' histories. We detected these domains by capturing all outgoing requests when a user visited a Web site. If there was at least one request from this site to a 3rd-party domain with Referer field in the HTTP header containing the site's URL, this domain is considered *being aware* of this visit. Table IV shows the top 25 trackers with their average percentage of tracked user's history that we detected in our real user dataset. We observed that

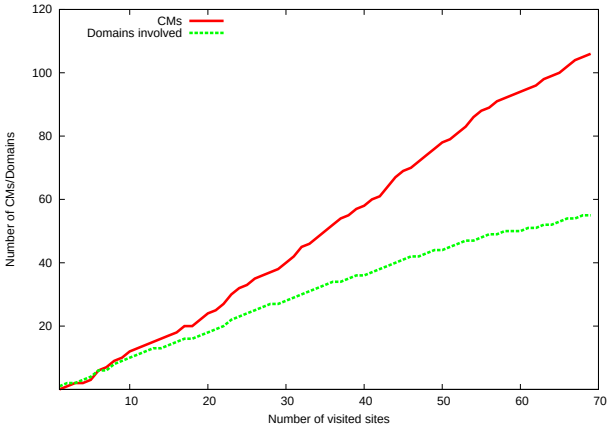


Fig. 5: Cookie matching frequency

TABLE III: Top pairs of domains executing cookie matching the most

Pair of domains	Frequency (% profiles)
facebook.com - adnxs.com	91
turn.com - admeld.com	87
doubleclick.net - rfihub.com	86
doubleclick.net - adnxs.com	85
doubleclick.net - mathtag.com	85
adnxs.com - admeld.com	84
doubleclick.net - turn.com	80
atdmt.com - bing.com	80
demdex.net - acxiom-online.com	79
doubleclick.net - yieldmanager.com	77
invitemedia.com - admeld.com	73
mathtag.com - admeld.com	71
doubleclick.net - invitemedia.com	71
doubleclick.net - amazon-adsystem.com	70
rubiconproject.com - rfihub.com	70
adnxs.com - amazon-adsystem.com	69
adnxs.com - rfihub.com	68
turn.com - p-td.com	67
turn.com - rubiconproject.com	65
mathtag.com - facebook.com	64

among the companies in the top 20 pairs of companies using Cookie Matching most frequently (Table III), 56% of them are in our list of 25 top trackers (Table IV). Meanwhile, 36% of these top trackers are in the top 20 pairs of companies most often performing Cookie Matching. These results show that, although Cookie Matching is used by numerous companies, the top trackers are often more involved than others.

We detected that some companies in the list of 25 top trackers can considerably increase the size of their users' profiles if cooperating. For example in our experiments, Facebook and AppNexus respectively tracked 31.55% and 17.4% of a users history on average, and they performed CM in 91% of the studied profiles. Their total Web history coverage would increase to 39.35%, on average, if they were merging their user histories. Table V shows some examples of the potential

TABLE IV: Top trackers

Tracker	Average (% of user history)
google-analytics.com	56.38
doubleclick.net	50.72
scorecardresearch.com	38.57
facebook.com	31.55
google.com	24.92
googleapis.com	23.84
facebook.net	23.44
quantserve.com	23.17
twitter.com	22.65
googleadservices.com	20.47
googlesyndication.com	20.41
2mdn.net	18.17
fbcdn.net	17.76
gstatic.com	17.56
adnxs.com	17.4
imrworldwide.com	15.73
yieldmanager.com	13.39
cloudfront.net	11.11
bluekai.com	10.92
atdmt.com	10.24
invitemedia.com	10.09
googletagservices.com	9.39
turn.com	9.21
rubiconproject.com	8.65
mathtag.com	8.01

TABLE V: Potential percentage of profile tracked after combination. Averages and ith quantiles.

Domains	Avg. (%)	Q_1 (%)	Q_2 (%)	Q_3 (%)
doubleclick.net - adnxs.com	52.43	48.74	52.86	56.34
doubleclick.net - yieldmanager.com	52.01	48.54	52.7	56.82
facebook.com - adnxs.com	39.35	36.0	39.47	44.3
adnxs.com - amazon-adsystem.com	19.32	16.05	18.67	22.35
invitemedia.com - admeld.com	14.12	11.84	14.29	16.44

combined profile sizes in cases of other companies. In this table, Q_1, Q_2, Q_3 are the first, second, and third quantiles respectively, computed among the 100 studied profiles.

A case study of Google and DoubleClick. Based on the results from Table IV, Google possesses 8 domains belonging to the top trackers: google-analytics.com (56.38%)[4], doubleclick.net (50.72%), google.com (24.92%), googleapis.com (23.84%), googleadservices.com (20.47%), googlesyndication.com (20.41%), gstatic.com (17.56%), googletagservices.com (9.39%). Although cookies used for these domains are all different, it is trivial to match them, for example by inspecting the IP address. By combining all data tracked by these domains Google could possibly know 80.13% of a user's visited sites, on average.

DoubleClick's Cookie Matching services are utilized by a substantial number of 3rd-parties. By analyzing Cookie Matching communications in all our experiments and tests[5], we extracted the host names of DoubleClick's Cookie Matching partners and counted their distinct top-level domain

[4]Even though Google Analytics uses unique cookies per sites, it is potentially possible to link these cookies across sites, for example by leveraging user's IP address

[5]Including all experiments we conducted in this work (section IV and V)

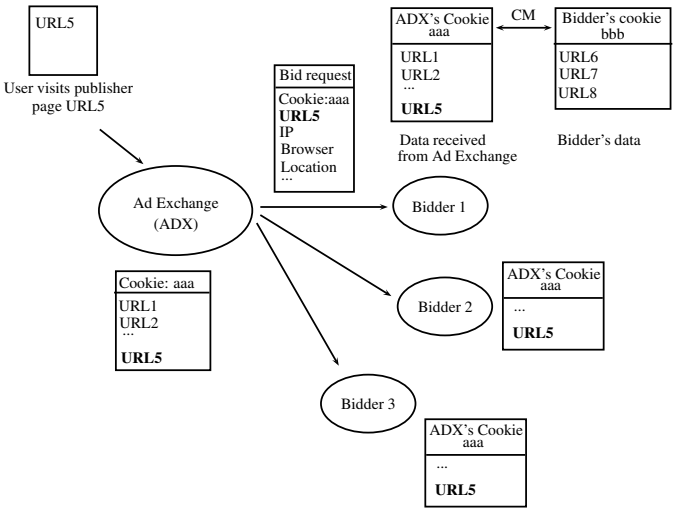


Fig. 6: Information leakage in Real-Time Bidding

names. For example, *dis.ny.us.criteo.com*, *dis.jp.as.criteo.com* and *dis.eu.criteo.com* share the same top-level domain name, *criteo.com*, and were counted once. In total, we detected 125 top-level domains performing CM with DoubleClick. It is interesting to note that one of the detected domain names was `e.visualdna.com` which belongs to a Big Data analytics company specializing in psychometrics, Visual DNA[6]. This example shows that Cookie Matching is not only used by advertisers, but also by other entities.

D. Real-Time Bidding privacy analysis

1) Privacy analysis: By combining RTB and CM, users' private data could potentially be leaked to bidders involved in real-time auctions. Figure 6 illustrates this leakage. We assume a situation between an Ad Exchange ADX holding the real-time auction, and a set of bidders $B_1, ..., B_k$ registering for the auction. Whenever a user visits a Web site W which requests ads from ADX, ADX sends a bid request to all the bidders. The bid request includes W's URL, the ADX's cookie of the user along with other additional information as discussed in section II-B. In our study we focus on the leaks of browsing histories, although it is evident that the additional information can potentially be used to fingerprint the user's browser [16]. Each time a bidder receives a bid request, he can save the W's URL and the ADX's cookie, resulting in a list of URLs assigned to each specific cookie from the ADX, provided that several bid requests containing this cookie were seen previously. Whether Cookie Matching takes place before or after these RTB processes, the bidder can combine all the previously-observed users' visited sites from received bid requests with its own user's profile identified by its own cookie. Even if Cookie Matching does not happen, these URLs can still provide a significant amount of information about the user identified by the ADX's cookie.

2) Methodology: We aim to show the frequency of RTB communications and quantify the information leakage described in the previous section. We examined all Real-Time Bidding requests, which we detected using our RTB detection technique (section IV-A), in our 100 profiles and extracted the related Ad Exchanges and winning bidders. The winning bidders were identified by the domain of each request, while the Ad Exchanges by the domain of the parent request in the request hierarchy (as discussed in III-A). We obtained a list of Ad Exchanges and for each Ad Exchange, a list of its bidders that won at least one auction. Examples of winning bidders in the case of DoubleClick Ad Exchange include AppNexus, AdRoll and InviteMedia.

We examined all profiles in the real user dataset. If a RTB event was detected on a site of a given profile, we assumed that all bidders participating in the RTB auction received the site's URL via the bid request sent by the Ad Exchange. We obtained the list of bidders associated to a given Ad Exchange by the use of methods described in the previous paragraph. We analyzed all sites in all the profiles and we counted how many sites would be leaking to each bidder via this mechanism. Subsequently, we divided the numbers of leaking sites by the total number of sites in the profile in order to quantify the history leakage.

It should be noted that each bidder can bid on several RTB Ad Exchanges, hence possibly learn parts of a user's browsing history from each of them. For example, we detected that AppNexus bids simultaneously on DoubleClick's and Admeld's RTB auctions.

We considered all the URLs that an Ad Exchange possibly sent to a bidder (detected by the above mechanism) as the *total leakage*. However, if a Cookie Matching event was detected between the Ad Exchange and the bidder during the experiment, we considered the URL leakage as a *matchable leakage*, otherwise *unmatchable leakage*. In matchable leakage, bidders can obviously combine profiles obtained from Ad Exchanges with their own users' profiles using Cookie Matching (Figure 6). Meanwhile, in *unmatchable leakage*, it is not clear whether the Cookie Matching will happen in the future, or other techniques can be used to link the two profiles. We therefore consider that the leakage is less severe in this case. *Total leakage* comprises both these two cases.

3) Results: Figure 7 shows the average cumulated number of RTB events, distinct Ad Exchanges and winning bidders after each visited site in our profiles. The cumulated numbers of RTB events after n visited sites are averaged from those numbers computed for each profile (red line). The average cumulated number of distinct Ad Exchanges and distinct winning bidders are shown in green and blue respectively. The figure shows that, when considering web histories of size 70, RTB occurred in 10% of the sites.

Figure 8 presents the percentage of user's history the three companies, Turn, AppNexus and InviteMedia, could obtain from Ad Exchanges in RTB. The figure shows a Complementary Cumulative Distribution Function (CCDF) of the percentage of user's history leak among the 100 profiles as well as their average (E) and standard deviation (D). The blue line represents the CCDF for the total leakage, while the red one represents the matchable leakage. The total leakage on average is around 11% of a user's history, but can be as high as 27% for certain profiles. With such high percentage of history received through RTB, even without Cookie Matching,

[6]http://www.visualdna.com

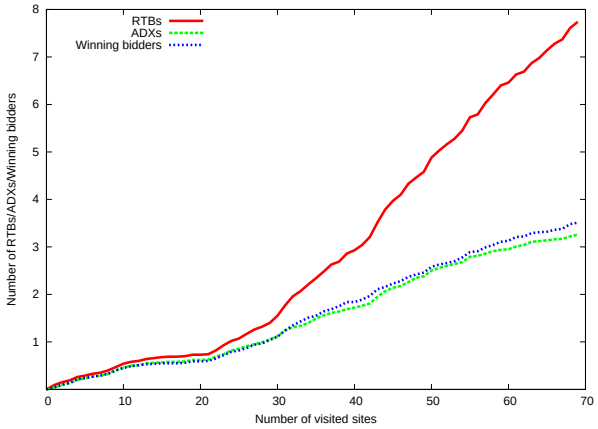


Fig. 7: Real-Time Bidding frequency

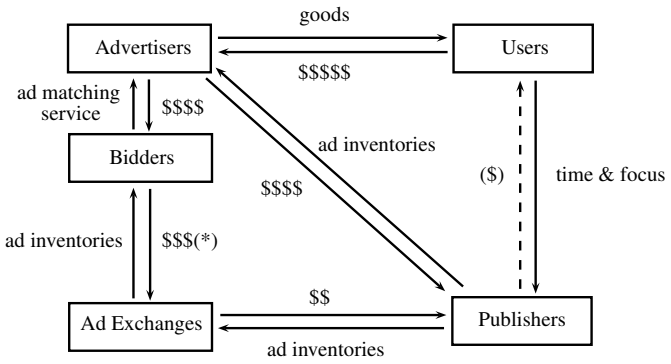


Fig. 9: Monetary flows in advertising systems. The communication we monitored is indicated by (*). Source: [45].

these companies can maintain a meaningful profile of a user. The matchable leakage is slightly lower, with the average value around 8% of the user's history. Bidders can easily combine these matchable data to their own users' profiles using Cookie Matching.

These numbers show that the user history leakage through RTB is significant. Given the fact that we only detected the lower bound of Real-Time Bidding communications, and that our assumption for the leakage is restricted to the bidders who won at least one observable auction, the leakage is potentially much higher in reality. Also, due to the rapid growth of RTB [26], these numbers are expected to considerably increase in the foreseeable future.

Information dissemination in RTB. We detected 41 winning bidders for all Ad Exchanges in total. In the case of DoubleClick Ad Exchange, we detected its 20 winning bidders, and 125 Cookie Matching partners which are likely real-time bidders as well. Although we did not encounter PulsePoint's Ad Exchange in all our experiments and tests[7], we found from its description a list of 59 RTB bidders [34]. These numbers suggest that 20-125 bidders might receive Web users' information in the case of DoubleClick, and at least 59 in the case of Pulse Point, which potentially constitutes a considerable information leakage.

V. Real-time price analysis

The observation of clear-text prices allows us to study how much advertisers pay for serving ads to users[8]. As discussed in section II-C, we believe that the prices paid in Real-Time Bidding reflect how bidders estimate the value of users. It is important to note that all prices reported in this section are represented in CPM (Cost-per-mille impressions), which means each price is for $1,000$ ad impressions. For example, a price of $\$0.12$ CPM or $\$0.12$ without any further explanation is actually $\$0.00012$ per impression.

[7]Including all experiments we conducted in this work (section IV and V)

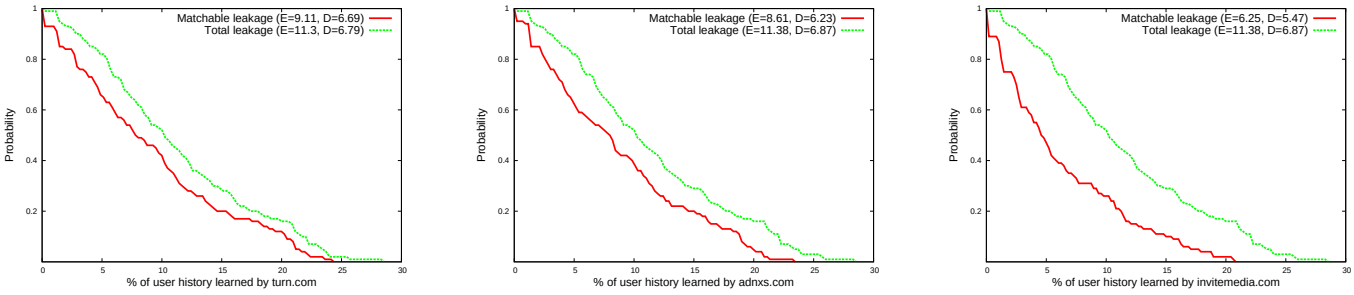
[8]Out of the $156,313$ prices we collected in the three countries, France, the US and Japan, $40,186$ were sent in clear-text (25.71%)

Figure 9 is a slightly modified version adopted from [45] and it shows the different monetary flows in a simplified model of advertising systems. The prices we retrieved for the analysis in this section are paid by *Bidders* to *Ad Exchanges* (marked with *(*)*).

Our analysis is based on clear-text prices we were able to detect. Despite the fact that we do not take into account encrypted prices, the prices we retrieved are comparable to the ones obtained directly from some Ad Exchanges' internal data and reported in other work [46], [45]. This constitutes a good evidence that encrypted and plain text prices are similar. In addition, we do not see any reason for advertisers to pay different prices on the basis of the price notification being encrypted or not.

Currency considerations. In our analysis, we assume that the currency used by different companies is USD. Our assumption is based on the fact that the majority of Ad Exchanges are US-based, which is also observable in our dataset, and that USD is the most commonly used currency in international business. Some Ad Exchanges, e.g. Pulse Point, publicly state the use of USD as the only currency in their RTB protocol description [35]. Since bidders/advertisers often reuse the URL patterns (the same domain and parameter names) to receive price notification from different Ad Exchanges, regardless of the price format, they are likely using the same currency. Finally, the value range of prices detected in our experiments is similar to those presented in other work leveraging internal advertisers' data [46], [27], which mention prices solely in USD.

Tracking consideration. Although there are many means of tracking the Web users, such as based on monitoring of IP addresses or fingerprinting techniques, cookie-based approach is still the dominant one. A good example is that Cookie Matching is a common technique used in RTB to match user profiles between two separate entities. Moreover during our experiments, we verified that targeted advertisements, for example ads about commercial products that users browsed previously, generally disappeared after clearing browser's cookies. In our work, we therefore assume that advertisers mostly rely on cookies to track users. Furthermore we assume that after clearing all the cookies of a browser, subsequent trackers perceive a request made by this browser as originating from a



(a) Turn (b) AppNexus (c) Invite Media

Fig. 8: CCDF of the percentage of user's history that bidders learned through RTB

new user.

We performed our analyses along the two angles that affect prices the most: (1) *Context analysis*: we study how the context, i.e. the category of the visited site, the user's physical location and time of visits, etc., influence prices, and 2) *Profile analysis*: we study how the user's profile, i.e. sites that the user have visited previously, affect prices.

A. Context analysis

Bid requests sent from Ad Exchange to its bidders contain specific context information about the user. These data might include the visited site's address, the user's physical location and timezone, and others; all these information are very likely impacting on advertisers' decisions. In this section, we aim to study this effect on bid prices.

1) Methodology: We developed a tool especially aimed at detecting and retrieving prices. Specifically, we re-implemented the core RTBAnalyzer plugin's functionality (section IV-A), keeping only the price retrieving function. We also developed scripts to automatically visit sites. We utilized, for this purpose, a WebKit browser PhantomJS[9] configured to act as a regular one. Our goal was to have a light-weight technique to capture prices with high performance.

2) Dataset: We investigated top sites from Alexa and kept only the $5,000$ first sites which contained RTB-capable scripts[10]. These scripts often come from URLs with known patterns, for example, pubads.g.doubleclick.net/gampad for DoubleClick [15]. Whenever we refer to a *sample*, we mean by this a HTTP request with a detected clear-text price.

3) Experiment Description: We subsequently visited each site in our $5,000$-site dataset. We applied a delay of 5 seconds between each two visits to ensure that the site was fully loaded, and in order to avoid being potentially blocked by sending too many ad requests in a short time. After visiting each site, we cleared cookies in order to ensure that advertisers had no prior information about the user in each visit. The $5,000$ sites were visited repeatedly (approximately 65 times each) during the month of June, 2013.

In order to study whether the physical location affects prices, we utilized Planet Lab [33] infrastructure to create dynamic tunnels to servers located in the US (New Haven) and Japan (Osaka). The French servers were located in Grenoble. The browser's timezone was set to local time for each of the analyzed countries. We ran the experiments simultaneously on these servers and used these three sets of collected prices to perform our Location, Time and Advertiser Analyses.

We aimed to minimize the potential effect of correlation between the studied aspects in our experiments and analyses. Specifically, with *Location*, we compared the results among the three datasets (i.e. US, Japan and France) given the same list of $5,000$ sites. With *Time* and *Advertiser* analyses, we studied the results within each dataset using the same list of visited sites. With *Site* and *Category* analyses, we examined the results from the same location (France). All experiments were designed such that they were distributed evenly by time during the day.

4) Results:

Site Analysis. For the used list of $5,000$ sites, we detected clear-text prices on $1,105$ of them. Figure 10 presents the average, minimum, maximum and median values of prices obtained by visiting these sites (X axis representing the site index), ordered by average. We only present data for the 630 sites for which we were able to collect at least 10 clear-text prices. As shown on the figure, the minimum and maximum prices differ wildly. The overall average price per site is $\$0.36$. Furthermore, average prices differ from site to site. For example, *ownersdirect.co.uk* has the average price of a mere $\$0.081$ (51 samples), while it is much larger in the case of *express.co.uk* at $\$0.98$ (65 samples). An interesting case is *officer.com* which has a remarkably large average price of $\$3.71$ (14 samples). These results suggest that some sites tend to be more "valuable" than others.

Category Analysis. We categorized the sites containing RTB ads with clear-text prices using Trend Micro [42]. We then grouped these sites by category and computed average prices for each category. Table VI shows the results for categories with number of prices (*Cnt* column) larger than

[9]http://phantomjs.org/

[10]We also analyzed top Alexa $5,000$ sites by visiting them 10 times and detected RTB ads on 467 of them. These sites were uniformly distributed in terms of Alexa ranking.

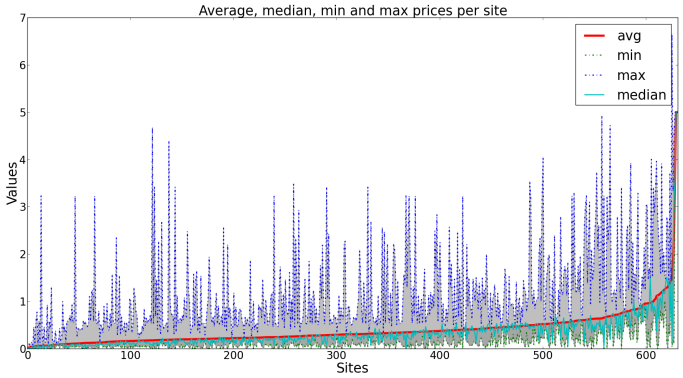


Fig. 10: Variation of prices per site. Data from France. Prices in CPM.

Category	Avg. price	Std	Median	Cnt	Pr%
Adult / Mature Content	0.25	0.15	0.22	515	1
Humor	0.25	0.19	0.20	710	1
Sports	0.29	0.18	0.36	1767	6
Games	0.32	0.16	0.42	2169	8
Blogs / Web Communications	0.33	0.25	0.32	2496	6
Entertainment	0.33	0.23	0.35	5005	15
Streaming Media / MP3	0.36	0.27	0.42	679	2
Computers / Internet	0.38	0.24	0.38	1450	6
News / Media	0.38	0.26	0.43	6913	23
Society / Lifestyle	0.38	0.27	0.46	707	3
Vehicles	0.41	0.34	0.37	643	3
Reference	0.48	0.21	0.61	577	2
Restaurants / Food	0.59	0.31	0.73	583	2
Shopping	0.68	0.38	1.10	633	2

TABLE VI: Average prices per category. Only categories with number of prices (Cnt) exceeding 500 taken into account. Data from France. Prices in CPM. Column Pr shows the percentage of sites belonging to a given category.

500[11], ordered by the average price. Among distinct categories, we observed several clear differences, indicating a category dependence. For example the average price for sites belonging to *Restaurants and Food* ($0.59) and *Shopping* ($0.68) are measurably larger than those associated with Humor ($0.25) and Blogs ($0.33). This suggests that visitors entering sites whose content belong to certain categories are much more worthy than visitors entering sites of other types.

Time Analysis. We grouped site visits by their time of execution into three 8-hour divisions of a day (0-8h, 8-16h and 16-24h); we used local times for each of the analyzed countries. Table VII presents the average price in each division. Highest prices were observed during the night and the early morning (0-8h). The trend was consistent in all three studied countries.

The results are similar to the ones in [46]: prices tend to be higher in the early morning. The authors argue that this is because there are more bidders competing over limited numbers of impressions in this time frame.

Location Analysis. The per-country averages are presented in Table VIII. The average price in the US is much higher than

[11]Consequently, the percentages of sites per category (*Pr* column) do not sum to 100.

Time division	The US	France	Japan
0-8h	0.75 (3246)	0.39 (10621)	0.28 (729)
8-16h	0.68 (2772)	0.36 (11375)	0.22 (732)
16-24h	0.62 (2520)	0.31 (7675)	0.19 (516)

TABLE VII: Average prices in different time divisions of day. Counts in parentheses. Prices in CPM.

Country	Average	Q_1	Q_2	Q_3	Count
The US	0.69	0.15	0.33	1.00	8538
France	0.36	0.11	0.24	0.47	29671
Japan	0.24	0.04	0.07	0.22	1977

TABLE VIII: Distribution of prices in three countries. Averages and ith quantiles. Prices in CPM.

Advertiser	US	FR	JP
mathtag.com	0.52 (862)	0.28 (4863)	0.30 (1303)
turn.com	0.65 (2659)	0.30 (7849)	0.06 (566)

TABLE IX: Average prices from different advertisers in three countries. Counts in parentheses. Prices in CPM.

in France, while Japan has the lowest average price. Table VIII also shows the first three quantiles of prices found in the US, France, and Japan.Most prices are really small, often less than $0.5 CPM.

Additionally, we examined advertisers detected in all considered locations and show the results in Table IX. The average prices in the case of `mathtag.com` in France and Japan are comparable, despite the fact that Japan has the lowest average compared to the other two countries. A common trend applies in the case of `turn.com`: average for Japan is very small. The average price in the US is still the highest in both these cases.

Advertiser Analysis. A separate Table X groups statistical data on prices for a subset of different advertisers (data for France) and shows that their bidding strategies differed. For example *adsrvr.org* bid much higher than *mathtag.com* did.

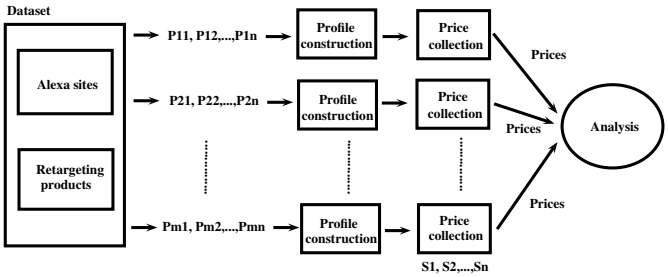
Furthermore, we detected clear-text price notifications in the case of DoubleClick as a bidder with a domain name bid. g.doubleclick.net and a price parameter `pr=`. DoubleClick's average price in France and in the US were $0.6 (102 samples) and $0.9 (38 samples) respectively. In both cases, DoubleClick's prices were much higher than average.

Given that truthful bidding is often encouraged in auctions, these results show that different entities possibly estimate users, or their private data, differently.

It is also interesting to note that the number of advertisers for which we observed clear-text prices varies among analyzed countries. Specifically, we detected 19, 8 and 6 such advertisers in the cases of the US, France and Japan respectively. Some of them are not very active, for example we encountered *rfihub.com* merely 17 and 2 times in the US and Japan respectively while some others are responsible for a large number of ads, with examples of *mathtag.com* and *turn.com* as shown in Table X.

Advertiser	Avg	Median	Stddev	Count
mathtag.com	0.28	0.09	0.44	4863
turn.com	0.30	0.19	0.30	7849
invitemedia.com	0.40	0.28	0.51	15481
adnxs.com	0.43	0.31	0.38	1242
doubleclick.net	0.60	0.20	0.72	102
adsrvr.org	0.63	0.56	0.41	102
w55c.net	0.84	0.62	0.76	30

TABLE X: Average prices from different advertisers. Data from France. Prices in CPM.



Pij is a site chosen from Alexa sites or retargeting products
S1, S2,..., Sk are the sites with prices. These sites are the same in all Price collection processes.

Fig. 11: Experiment for artificial profile analysis

B. Profile analysis

In this section, we study whether the user's Web browsing history (or *profile*) affects prices that advertisers pay for serving ads. Our methodology is to create a number of artificial profiles, use them to visit the same set of Web sites, collect and analyse winning prices.

Figure 11 summarizes our experiment. We built profiles considering two aspects: *history categories* (i.e. categories of visited sites)[12] and *intents* (e.g. browsing for a commercial product).

1) Methodology: In order to visit sites, we used Selenium [39] to instrument a Firefox browser equipped with RTBAnalyzer plug-in (described in section IV-A).

2) Dataset: We crawled 50 top Alexa sites from each of the following Alexa categories: *Adult, Arts, Business, Business-Financial Services, Computers, Games, Health, Home, Kids and Teens, News, Recreation, Science, Shopping, Sports*. Those sites were used to construct history categories in each profile.

For the intent construction, we used three commercial Web sites that are very popular in our country of residence (France): fnac.com (electronic products), hotels.fr (hotel booking) and maty.com (jewelry). We call them *retargeting sites* hereafter, as after visiting them, the previously-browsed products from these sites often appear in online ads during regular browsing. We randomly chose 5 products from each of these sites and kept these three lists of products to build users' intents.

From the dataset used in context analysis (section V-A2) we extracted a list of sites which often resulted with ads containing clear-text prices[13]. Among these sites, we extracted the top 17 sites which had the highest rate of clear-text price occurrence[14]. We call them *sites with prices*.

3) Experiment Description: We created 14 profiles for each of the following types:

- *New user*: empty profile
- *Only category*: only visit Alexa sites belonging to one category
- *Category + fnac.com*: visit Alexa sites belonging to one category, then visit 5 products on fnac.com
- *Category + hotels.fr*: visit Alexa sites belonging to one category, then visit 5 products on hotels.fr
- *Category + maty.com*: visit Alexa sites belonging to one category, then visit 5 products on maty.com
- *Only maty.com*: only visit 5 products on maty.com

We simultaneously ran 6 instances of Firefox browser to perform the tests with these profiles. Each instance was devoted to one kind of profile. For each profile in each browser's instance we subsequently performed a profile construction and price collection, and repeated this phase 10 times. All price collection processes were performed using the same set of *sites with prices*. We executed our experiments evenly throughout the day to ensure that time of day did not affect prices.

4) Results: We obtained 20 prices per profile and per round, consequently about 200 prices per profile in total (after 10 rounds), on average. The detected average prices per profiles are shown in Table XI. The prices for profiles "*Only category*" are about 40% higher than those for "*New user*". Among "*Only category*" profiles, different profile categories result in different prices. This is particularly acute for the category *Games*, which exhibits prices 38% higher than average. Other category profiles with prices larger than the average price are *Sports*, *Health*, and *Kids and Teens*. Our results show that the type of visited sites is actually affecting prices that advertisers paid for serving ads to users.

The results indicate that retargeted ads (the ones which match users' intents) often receive higher prices than those for "*Only category*". These prices also differed among different retargeting advertisers. For example, prices from fnac.com were the lowest, with average $0.64, prices related to hotels.fr were slightly higher with $0.69, whereas maty.com had the remarkably highest average price of around $1.2. This could be explained by the strategies of the different advertisers and possibly by the prices of the advertised products. Interestingly, we also noticed that maty.com retargeted ads were displayed much more frequently than fnac.com or hotels.fr ads. Finally, we observed negligible differences in prices between "*Only maty.com*" and "*Category and maty.com*". This clearly shows that even though users' browsing histories are taken into account when advertising a product, advertisers actually value users' intentions much more.

Although we expected that retargeted ads are related to higher prices, the striking difference between winning prices for ads after visiting maty.com's products and those after

[12] History categories can be used, for example in Google's and Yahoo's systems, to personalize ads [23][44].

[13] A sample of this list is available at http://yourvalue.inrialpes.fr.

[14] Examples of such sites are accuweather.com, tinyurl.com or technorati.com

TABLE XI: Artificial profile analysis. Prices in CPM.

Category	New user		Only category		Category + fnac.com		Category + hotels.fr		Category + maty.com		Only maty.com	
	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
Adult	N/A	N/A	0.44	0.20	0.56	0.27	0.64	0.20	1.12	0.21	N/A	N/A
Arts	N/A	N/A	0.51	0.17	0.52	0.15	0.66	0.23	1.28	0.29	N/A	N/A
Business	N/A	N/A	0.55	0.22	0.63	0.21	0.61	0.21	1.10	0.34	N/A	N/A
Business - Finan. Serv.	N/A	N/A	0.59	0.20	0.68	0.24	0.88	0.28	1.31	0.31	N/A	N/A
Computers	N/A	N/A	0.48	0.21	0.57	0.20	0.70	0.25	1.18	0.14	N/A	N/A
Games	N/A	N/A	0.80	0.35	0.74	0.29	0.81	0.40	1.41	0.27	N/A	N/A
Health	N/A	N/A	0.67	0.47	0.68	0.34	0.81	0.43	1.21	0.30	N/A	N/A
Home	N/A	N/A	0.58	0.21	0.70	0.39	0.57	0.23	1.00	0.23	N/A	N/A
Kids and Teens	N/A	N/A	0.64	0.33	0.65	0.27	0.74	0.29	1.25	0.27	N/A	N/A
News	N/A	N/A	0.50	0.12	0.72	0.38	0.74	0.29	1.09	0.18	N/A	N/A
Recreation	N/A	N/A	0.55	0.21	0.64	0.32	0.69	0.16	1.12	0.22	N/A	N/A
Science	N/A	N/A	0.50	0.19	0.60	0.37	0.59	0.21	1.36	0.24	N/A	N/A
Shopping	N/A	N/A	0.53	0.22	0.61	0.27	0.65	0.25	1.21	0.23	N/A	N/A
Sports	N/A	N/A	0.71	0.47	0.59	0.29	0.62	0.17	1.17	0.21	N/A	N/A
Average	0.41	0.10	0.58	0.26	0.64	0.29	0.69	0.26	1.20	0.25	1.17	0.26

visiting non-retargeting sites deserves a detailed analysis. To our knowledge, most of ad auctions apply the *second-price principle* which means the winning bidder only pays a slightly higher price than the second highest bid price[15]. In other words, the paid price is the second highest price incremented by a small value defined by the RTB. Assuming that the average winning price of a normal ad is $\$0.4$ and a retargeted ad has a significantly higher bid price of $\$1.2$, the average winning price should still be close to $\$0.4$. However, we observed much higher prices in the case of retargeted ads, specifically in the case of advertisements from maty.com, when served by Criteo. A possible explanation is that there are several competing retargeters who bid for this ad impression. In order to verify this, we conducted the following experiment: We used Ghostery [2] to block all other trackers except Criteo and a selected number of RTB systems (Admeld, AppNexus, Pubmatic and Rubicon; Criteo bids in their auctions). Similarly to the previous experiment, we browsed 5 products from maty.com and then a list of *sites with prices*. As described above, we blocked most of the trackers while browsing products on maty.com. We then stopped blocking trackers when visiting *sites with prices* when we aimed to detect the clear-text prices of advertisements. We performed this experiment 10 times. The average price observed during this experiment was $\$0.44$ CPM, much lower than previously when the average was $\$1.17$ (Table XI). In this setting Criteo could still win the auctions but at a much smaller cost (because we intentionally blocked the competitors). This result proved that other bidders had been involved in the initial scenario, and that retargeting companies are also competing on retargeted ads.

Advertisers also bid on new users. This could give them an opportunity to perform a Cookie Matching on them. As described in section II-B, the winner has the right to initialize a Cookie Matching with the Ad Exchange when serving ads through Real-Time Bidding mechanism. The price $\$0.0004$ per impression (average value $\$0.41$ CPM divided by 1000) would be very reasonable for the opportunity to track a new user.

[15]Note that RTB systems can fine-tune their internal auction parameters to effectively switch from second-price to first-price auctions [46]. However, as we show, this does not apply to our case.

TABLE XII: Real profile analysis. Prices in CPM.

Property	Value
Number of profiles with clear-text prices	89
Avg. number of prices per profile	3.83
Average price per profile	0.43
Standard deviation (price per profile)	0.37
Min price per profile	0.04
Max price per profile	1.98

The relatively large variance values in the results can be explained by the fact that ads prices depend on several parameters such as different campaigns or different bidders. Furthermore, RTB is by definition dynamic, thus consequently auctions could possibly be won by different advertisers in each round of our tests. It is also important to note that the variance for *new users* is much (2.6 times) lower than in the other cases.

5) Real profile analysis: We analyzed clear-text prices obtained in the real user dataset (section IV-B). The results are shown in Table XII. Among the 100 users, 89 had at least one clear-text price. The average number of clear-text prices per profile is approximately 4. There is a high rate of variation among the prices per analyzed profile, with minimum at $\$0.04$, maximum at $\$1.98$, and average value of $\$0.43$. We also investigated the 8 profiles which had at least 7 prices per profile to analyze how prices vary among them. Table XIII presents the number (count), average value and standard deviation of prices in 8 profiles from our dataset. The prices we observed with real user profiles actually vary within the value range of prices detected in artificial profiles as shown in Table XI.

VI. Discussion

A. Data exchange between companies

Data exchange is a growing trend in modern advertising systems. When targeted ads become increasingly sophisticated, the users' dataset maintained by an intermediary (e.g. an ad network) might not adequately meet these demands. Naturally, advertisers desire to target users with the use of their own

TABLE XIII: Real profile examples. Prices in CPM.

Index	Average price	Standard deviation	Count
1	0.16	0.17	7
2	0.26	0.14	11
3	0.41	0.51	8
4	0.43	0.20	8
5	0.45	0.31	8
6	0.91	0.68	13
7	1.11	0.89	7
8	1.13	1.00	8

data as well. For example, Facebook has been working with data vendors Datalogix, Epsilon, Acxiom and BlueKai [17] in order to allow its clients to serve ads based on their offline data [11]. RTB services enable advertising companies to use their own online data for serving targeted ads. While this data exchange is expected to enhance advertising performance, it should be designed with careful consideration. Otherwise, this could lead to users' data leakage between various companies and the resulting loss of control over this data. In this paper, we showed that this might indeed be the case with RTB. We investigated and quantified the leakage based on the assumption of non-adversary parties. With malicious attempts, e.g. collusion between the companies with the aim to combine their users' profiles, the risks could be much more severe.

B. Privacy-preserving targeted advertising system

There have been a considerable number of research work towards designing a targeted advertising system not utilizing tracking, such as Privad [38] and Adnostic [41]. Yet, most of the proposed solutions are designed in the traditional ad network setting. Their common idea is to save users' profiles on the client side; ad networks send coarse-grained ads to the client, which then can locally select the most appropriate ones to display, according to the user profile. It is not clear if these systems can be adapted to new technologies such as RTB. In RTB, the advertisers want to customize their bids towards each individual user, e.g. a jewelry advertisement could have different values when showing to a male and a female. Moreover, advertisers are likely interested in adjusting their buying strategy at real time. The emergence of such new demands and techniques requires a significant change in the proposed privacy-preserving targeted advertising systems, or even a new design approach, in order to address privacy problems while maintaining current business models.

C. The economics of private data

In a study performed by Carrascal et al [7], users evaluate the disclosure price of their presence on a Web site to EUR 7, on average. In this work, we showed that this information is actually being sold off at a much lower price by Ad Exchanges and that its price depends on the user's browsing profile, in addition to other contextual information. Our experiments demonstrated that, on average, the presence of a user on a Web site is sold to the winner of the RTB auction for less than $\$0.0005$ ($\$0.5$ CPM). We also note that since the presence of the user on a Web site is actually broadcast to all the bidders during a RTB request, this cost can be shared among them. The actual cost per bidder could then be computed by dividing $\$0.0005$ by the number of bidders, which we estimated to $20 - 125$ for DoubleClick. We acknowledge that the cost also includes the price paid for the ad delivery. The huge gap between these figures and those from the users' perception can be explained by the fact that user information is currently extremely easy to collect (e.g. by simply placing a small JavaScript code in a Web site), therefore could be sold at very cheap price.

Revenue per user. Estimating how much advertisers spend on a user is an interesting problem, and we aim to provide a rough estimation of this cost. According to the work of Castelluccia et al. [9], targeted ads account for about 30% of total ads. From the analyses in the previous section, we can assume that the average price per ad is $\$0.0005$.

We manually counted advertisements on 50 sites corresponding to an one-day browsing history of a volunteer and detected 40 ads in total (0.8 ad per site). We therefore derived the total number for targeted ads at around 12 (30% of ads on 50 sites) in the analyzed case. Setting the average price per ad to $\$0.0005$, these ads cost advertisers $\$0.006$ per day. Accordingly, the cost is approximately $\$0.18$ per month and $\$2.16$ per year. If, for example, Ad Exchanges take a commission fee of 20% for each transaction, they could earn around $\$0.432$, and the publishers $\$1.728$, per user, per year.

This simplified scenario is only meant as a rough estimation since many aspects remain uncertain. For example, the number of ads per site and the number of browsed sites per day may not be representative. We also assumed that all other cost models such as Cost Per Click (CPC) or Cost Per Action (CPA) can be converted to the equivalent CPM. For example a price of $\$0.01$ CPC for an ad with click probability of 10% can be converted to a $\$1$ CPM. We therefore assumed in our estimation that the average CPM price (established in previous sections) applies to all targeted ads. By this estimation, we showed an initial quantification of how much a user costs, or how much money which entities (Ad Exchanges, publishers, etc.) gain from the user's data in online advertising market.

VII. Related work

User tracking and resulting privacy risks have been discussed in a plethora of research studies, notably [28], [29], [30], [32], [8], [9]. These work primarily showed the sensitiveness of user's data and the different possibilities of users' Web browsing history leakage. For example, Web history can be leaked through Web search suggestion [8], or targeted ads [9]. Furthermore, a Web history itself can be used to fingerprint users [32]. In this paper, we leveraged another privacy leakage channel in RTB, which potentially allows companies to significantly increase their tracking horizon.

Roesner and Kohno [37] proposed a taxonomy to classify tracking behaviors beyond the simple notions of first- and third-party tracking. The information leakage in RTB, presented in this paper, falls in the tracking *Behavior D* category from their classification framework, i.e. information leakage happens through an intermediate party. Nevertheless, while Roesner and Kohno only considered tracking activities that are visible on the client side, RTB leakage can be considered *invisible*, as it happens on the server side and therefore can hardly be detected by existing methodologies. Moreover, we

not only described the leakage, but also quantified the amount of leaked information. It is also worth pointing out that RTB is a data exchanging hub; users' data are being transferred during auctions.

Yuan et al. [46] described RTB and Ad Exchange mechanism and provided an in-depth analysis of a production Ad Exchange. Based on the internal auction data of the Ad Exchange, they built time-dependent models of bid prices. Interestingly, they showed that a publisher's soft floor configuration can change from second-price to first-price auction, and therefore cause losses to advertisers. In summary, their conclusion is that the current bidding strategy is far less optimal, requiring optimization algorithms. In some respects, they provided similar findings to ours, for example, that prices tend to be higher in the early morning. However, while they mostly focused on the economic aspect, we paid attention to user's privacy. Specifically, we also studied how advertisers customize prices according to different profiles of users.

The value of user's private data has long been an interesting topic and attracted a considerable body of research [5], [7], [10], [40]. We categorize these work into two main approaches: from users' [5], [7], [10] and advertisers' [40] perspectives. The work presented in Financial Times [40] provided an analysis of industry pricing data from a range of sources in the US. The authors showed that general personal information, such as age, gender and location is worth a mere $\$0.0005$. A person who is having a specific intent, e.g. buying a car, is likely worth more at about $\$0.0021$. Although the used data source and methodology are not published, the results are quite similar to ours. From the user's side, Danezis et al. [10] analyzed how users evaluate their location data, or in other words, the compensation they expect to receive for making their location data available to advertisers. Acquisti et al. [5] discussed the value of privacy according to the two concepts: *Willingness To Pay* (the monetary amount users are willing to pay to protect their privacy) and *Willingness To Accept* (the compensation that users are willing to accept for their privacy loss). Carrascal et al. [7] provided the most specific results from the users' perspective, showing that on average, users evaluate the price of the disclosure of their presence on a Web site in their browsing history to EUR 7. We showed that this information is actually sold off at a much lower price by Ad Exchanges (section VI-C). In summary, the methodologies of [5], [7] and [10] are mostly based on user surveys, while our approach is entirely different and purely empirical.

In a broader perspective, a number of research work studied the impact of users' data in online advertising from the economic angle. Interestingly, [18] showed that with a simple and common mechanism (second price auction with a reserve price), incorporating user's data might decrease the revenue of auctioneers. Mahdian et al. [31] discussed another interesting economic aspect of Cookie Matching: whether premium publishers loose their revenues as a result of Cookie Matching, since advertisers might follow their users and show ads to them in other non-premium publishers' Web sites with a lower cost. Nevertheless, they concluded that this is not the case; when advertisers are homogeneous, the publishers agree about the benefit of Cookie Matching: either they all benefit or suffer from it. The work of Johnson [27] studied the impact of different possible privacy policies. Its results suggest that online publisher revenues drop by 3.9% under an opt-out policy, 34.6% under an opt-in policy, and 38.5% under a tracking ban. Total advertiser surplus drops by 4.6%, 40.9%, and 45.5% respectively.

VIII. Conclusion

In this work, we characterized Real-Time Bidding (RTB) and Cookie Matching (CM), and highlighted the core privacy risks associated with the use of these technologies. We showed that RTB and CM are observably prevalent on the Web and lead to significant user information leakage. Concretely, RTB can leak as much as 27% of a user's Web browsing history to a bidder involved in Ad Exchanges' auctions. The actual leakage is expected to be higher, since we only established a lower bound of actual RTB communications. The process is inherently non-transparent, and this *invisible* leakage cannot be observed using current tracking measurement tools such as Collusion [1] and Ghostery [2]. Nevertheless, a strict privacy protection approach, such as blocking all ad-related URLs using Ghostery or AdBlock Plus [6] could potentially solve this privacy problem.

RTB creates a data market where users' browsing data are sold at auctions to advertisers. We showed that advertisers are evaluating each individual user differently depending on several criteria. Our results indicate that the presence of a user in a Web site is often sold off for less than $\$0.0005$, which is far lower than that from users' perception [7]. We highlight that such sophisticated methodologies being used to commoditize users data without their awareness, let alone consent, is a problem that needs due attention.

References

[1] Collusion. https://www.mozilla.org/en-US/collusion/.

[2] Ghostery. http://www.ghostery.com/.

[3] Httpfox. https://addons.mozilla.org/en-US/firefox/addon/httpfox/.

[4] A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. Security & Privacy, IEEE, 3(1):26–33, 2005.

[5] A. Acquisti, L. John, and G. Loewenstein. What is privacy worth. In Workshop on Information Systems and Economics (WISE), 2009.

[6] Adblock Plus. Adblock plus - surf the web without annoying ads! https://adblockplus.org.

[7] J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira. Your browsing behavior for a big mac: Economics of personal information online. arXiv preprint arXiv:1112.6098, 2011.

[8] C. Castelluccia, E. De Cristofaro, and D. Perito. Private information disclosure from web searches. In PETS, 2010.

[9] C. Castelluccia, M.-A. Kaafar, and M.-D. Tran. Betrayed by your ads! In PETS, 2012.

[10] G. Danezis, S. Lewis, and R. J. Anderson. How much is location privacy worth? In WEIS. Citeseer, 2005.

[11] C. Delo. Facebook to partner with acxiom, epsilon to match store purchases with user profiles. http://adage.com/article/digital/facebook-partner-acxiom-epsilon-match-store-purchases-user-profiles/239967/.

[12] DoubleClick. Doubleclick ad exchange real-time bidding protocol. https://developers.google.com/ad-exchange/rtb/.

[13] DoubleClick. Processing the request – example bid request. https://developers.google.com/ad-exchange/rtb/request-guide\#example-bid-request.

[14] DoubleClick. Real-time bidding protocol request examples. https://developers.google.com/ad-exchange/rtb/downloads/realtime-bidding-proto.txt.

[15] DoubleClick Help. Serve ads in a non-javascript environment. https://support.google.com/dfp_premium/answer/1638620?hl=en.

[16] P. Eckersley. How unique is your web browser? In Privacy Enhancing Technologies, pages 1–18, 2010.

[17] Facebook. Updates to custom audiences targeting tool. http://newsroom.fb.com/News/576/Updates-to-Custom-Audiences-Targeting-Tool.

[18] H. Fu, P. Jordan, M. Mahdian, U. Nadav, I. Talgam-Cohen, and S. Vassilvitskii. Ad auctions with data. In Algorithmic Game Theory, pages 168–179. Springer, 2012.

[19] Google. The arrival of real-time bidding. http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/pl//doubleclick/pdfs/Google-White-Paper-The-Arrival-of-Real-Time-Bidding-July-2011.pdf.

[20] Google. Cost-per-click (cpc). https://support.google.com/adsense/answer/18196?hl=en.

[21] Google. Cpm ads. https://support.google.com/adsense/answer/32725?hl=en.

[22] Google. Decrypting price confirmations. https://developers.google.com/ad-exchange/rtb/response-guide/decrypt-price.

[23] Google. Google ads settings. https://www.google.com/settings/u/0/ads.

[24] Google. Google's cookie matching protocol. https://developers.google.com/ad-exchange/rtb/cookie-guide.

[25] IAB. Openrtb api specification version 2.1. http://openrtb.googlecode.com/files/OpenRTB-API-Specification-Version-2-1-FINAL.pdf.

[26] IDC. Real-time bidding in the united states and western europe, 20102015. http://info.pubmatic.com/rs/pubmatic/images/IDC_Real-Time%20Bidding_US_Western%20Europe_Oct2011.pdf.

[27] G. Johnson. The impact of privacy policy on the auction market for online display advertising. http://gradstudents.wcas.northwestern.edu/~gaj741/GarrettJohnson-JMP.pdf.

[28] B. Krishnamurthy, D. Malandrino, and C. E. Wills. Measuring privacy loss and the impact of privacy protection in web browsing. In Proceedings of the 3rd symposium on Usable privacy and security, pages 52–63. ACM, 2007.

[29] B. Krishnamurthy, K. Naryshkin, and C. Wills. Privacy leakage vs. protection measures: the growing disconnect. In Web 2.0 Security and Privacy Workshop, 2011.

[30] B. Krishnamurthy and C. Wills. Privacy diffusion on the web: a longitudinal perspective. In Proceedings of the 18th international conference on World wide web, pages 541–550. ACM, 2009.

[31] M. Mahdian, A. Ghosh, P. McAfee, and S. Vassilvitskii. To match or not to match: Economics of cookie matching in online advertising. In Proceedings of the 13th ACM Conference on Electronic Commerce, pages 741–753. ACM, 2012.

[32] L. Olejnik, C. Castelluccia, A. Janc, et al. Why johnny can't browse in peace: On the uniqueness of web browsing history patterns. In 5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012), 2012.

[33] Planet Lab. Planet lab – an open platform for developing, deploying, and accessing planetary-scale services. http://www.planet-lab.org/.

[34] Pulse Point. Private exchange. http://docs.pulsepoint.com/display/RTB/Private+Exchange.

[35] PulsePoint. Pulsepoint real-time bidding api. http://docs.pulsepoint.com/display/RTB/Real-Time+Bidding+API.

[36] PulsePoint. Real-time bidding protocol request examples. http://docs.pulsepoint.com/display/RTB/Real-Time+Bidding+API#Real-TimeBiddingAPI-BidRequestParameters.

[37] F. Roesner, T. Kohno, and D. 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, pages 12–12, Berkeley, CA, USA, 2012. USENIX Association.

[38] B. C. Saikat Guha and P. Francis. Privad: practical privacy in online advertising. In NSDI, 2011.

[39] Selenium. Selenium - web browser automation. http://docs.seleniumhq.org/.

[40] E. Steel, C. Locke, E. Cadman, and B. Freese. How much is your personal data worth? http://www.ft.com/intl/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html.

[41] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas. Adnostic: Privacy preserving targeted advertising. In NDSS, 2010.

[42] Trendmicro. Trend micro site safety center. http://global.sitesafety.trendmicro.com.

[43] W. Vickrey. Counterspeculation, auctions, and competitive sealed tenders. The Journal of finance, 16(1):8–37, 1961.

[44] Yahoo. Ad interest manager. http://info.yahoo.com/privacy/us/yahoo/opt_out/targeting/.

[45] S. Yuan, A. Z. Abidin, M. Sloan, and J. Wang. Internet advertising: An interplay among advertisers, online publishers, ad exchanges and web users. arXiv preprint arXiv:1206.1754, 2012.

[46] S. Yuan, J. Wang, and X. Zhao. Real-time bidding for online advertising: Measurement and analysis. arXiv preprint arXiv:1306.6542, 2013.