If you are not paying for it, you are the product: How much do advertisers pay to reach you?

Panagiotis Papadopoulos
FORTH-ICS, Greece
panpap@ics.forth.gr

Nicolas Kourtellis
Telefonica Research, Spain
nicolas.kourtellis@telefonica.com

Pablo Rodriguez Rodriguez
Telefonica Alpha, Spain
pablo.rodriguezrodriguez@telefonica.com

Nikolaos Laoutaris
Data Transparency Lab, Spain
nikos@datatransparencylab.org

ABSTRACT

Online advertising is progressively moving towards a programmatic model in which ads are matched to actual interests of individuals collected as they browse the web. Letting the huge debate around privacy aside, a very important question in this area, for which little is known, is: *How much do advertisers pay to reach an individual?*

In this study, we develop a first of its kind methodology for computing exactly that – the price paid for a web user by the ad ecosystem – and we do that in real time. Our approach is based on tapping on the Real Time Bidding (RTB) protocol to collect cleartext and encrypted prices for winning bids paid by advertisers in order to place targeted ads. Our main technical contribution is a method for tallying winning bids even when they are encrypted. We achieve this by training a model using as ground truth prices obtained by running our own "probe" ad-campaigns. We design our methodology through a browser extension and a back-end server that provides it with fresh models for encrypted bids. We validate our methodology using a one year long trace of 1600 mobile users and demonstrate that it can estimate a user's advertising worth with more than 82% accuracy.

CCS CONCEPTS

•Information systems →Online advertising; Display advertising; *Web log analysis;*

KEYWORDS

Ad Transparency, Cost of Advertising, RTB Auctions, Pricing Dynamics, User Privacy

ACM Reference format:
Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. 2017. If you are not paying for it, you are the product:
How much do advertisers pay to reach you?. In *Proceedings of IMC '17, London, United Kingdom, November 1–3, 2017,* 15 pages.
DOI: 10.1145/3131365.3131397

IMC '17, London, United Kingdom
© 2017 ACM. 978-1-4503-5118-8/17/11…$15.00
DOI: 10.1145/3131365.3131397

1 INTRODUCTION

In today's data-driven economy, the amount of user data an IT company holds has a direct and non-trivial contribution to its overall market valuation [73]. Digital advertising is the most important means of monetizing such user data. It grew to $194.6 billion in 2016 [75] of which $108 billion were due to mobile advertising. In fact, more and more companies rush to participate in this rapidly growing advertising business either as advertisers, ad-exchanges (ADXs), demand-side platforms (DSPs), data management platforms (DMPs), or all of the above. For these companies to increase their market share, they need to deliver more effective and highly targeted advertisements. A way to achieve this is through programmatic instantaneous auctions. An important enabler for this kind of auctions is the Real-Time Bidding (RTB) protocol for transacting digital display ads in real time. RTB has been growing with an annual rate of 128% [80], and currently accounts for 74% of programmatically purchased advertising. In US alone it created a revenue of $8.7 billion in 2016 [8].

Consequently, the collection of user personal data has become more aggressive and sometimes even intrusive [29, 33], raising a huge public debate around the tradeoffs between (i) innovation in advertising and marketing, and (ii) basic civil rights regarding privacy and personal data protection [51, 55]. These increasing privacy concerns, drew the attention of a significant body of research, which studied users' privacy loss in conjunction to existing user tracking techniques [1, 17, 21, 52, 60], and proposed various defence mechanisms to users [59, 64, 65]. Still, there is an outstanding question that remains unaddressed by the related work in the area. This question concerns transparency and is the following: *Based on the exposed user personal data, how much do advertisers pay to reach an individual?*

Despite the importance of this question, it is surprising how little is known about it. There exist several reports about the *average* revenue per user (ARPU) from online advertising [13, 30, 67], but ARPU, as its name suggests, is just an average. It can be calculated by dividing the total revenue of a company by the number of its monthly active users. Computing the revenue per *individual* user is a completely different matter for which very limited work is available.

In particular, the FDVT [14] browser extension can estimate the value of an individual user for Facebook, by tapping on the platform's ad-planner. Another important prior work [62] leverages similarly the RTB protocol and specifically its final stage, where the winning bidder (advertiser) gets notified about the auction's charge price per delivered impression. These charge prices were initially transmitted in cleartext and focused solely on them. However, more and more advertising companies use encryption to reduce the risk of tampering, falsification or monitoring from competitors. This trend renders that method inapplicable for the current and future ad ecosystem, whose majority of companies will deploy charge price encryption. In contrast to these works, our present method takes into account all the web activity of a user (not only on Facebook), and all RTB traffic, i.e., both cleartext and encrypted prices.

In this paper, our motivation is to enhance *transparency* in digital advertising and shed light on pricing dynamics in its personal data-driven ecosystem. Therefore, we develop and evaluate a first of its kind methodology for enabling end-users to estimate in real time their actual cost for advertisers, even when the latter encrypt the prices they pay. Designed as a browser extension, our method can tally winning bids for ads shown to a user and display the resulting amount as she moves from site to site in real time.

In summary, we make the following main contributions:

(1) We propose the first to our knowledge holistic methodology to calculate the overall cost of a user for the RTB ad ecosystem, using both encrypted and cleartext price notifications from RTB-based auctions.

(2) We study the feasibility and efficiency of our proposed method by analyzing a year-long weblog of 1600 real mobile users. Additionally, we design and perform an affordable (a few hundred dollars cost) 2-phase real world ad-campaign targeting ad-exchanges delivering cleartext and encrypted prices in order to enhance the real-users' extracted prices. We show that even with a handful of features extracted from the ad-campaign, our methodology achieves an accuracy > 82%. The resulting ARPU is ~55% higher than that computed based on cleartext RTB prices alone. Our findings challenge the related work's basic assumption [62] that encrypted and plain text prices are similar (we found encrypted prices to be ~1.7× higher). Finally, we validate our methodology by comparing our average estimated user cost with the reported per user revenue of major advertising companies.

(3) Using lessons from the study, we design a system where the users, by using a Chrome browser extension, can estimate in real-time, in a privacy-preserving fashion on the client side, the overall cost advertisers pay for them based on their exposed personal information. In addition, they can also contribute anonymously their impression charge prices to a centralized platform for further research.

Paper Organization. Section 2 summarizes various key concepts of the RTB ecosystem and presents the main challenge and motivation of our work. Section 3 provides a high-level overview of our novel methodology and our price modeling engine. Section 4 presents an analysis of the dataset we use to bootstrap our modeling of encrypted prices. Section 5 presents in detail the effort to model RTB charge prices by executing probing ad-campaigns. These campaigns provide ground-truth data, which is used to train a machine learning classifier that can estimate encrypted prices in real-time at the browser of a user. Section 6 puts all the pieces together and presents results on the overall monetary cost for displaying ads to users. Section 7 covers related work while Section 8 discuses various aspects of our work and concludes the paper.

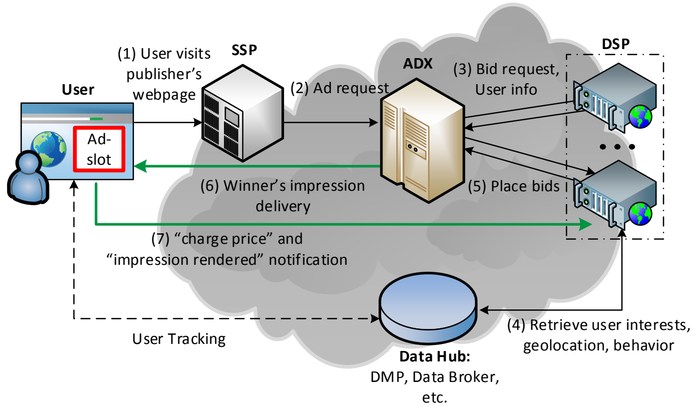

User
Ad-slot
(1) User visits publisher's webpage
SSP
(2) Ad request
ADX
(3) Bid request, User info
DSP
(5) Place bids
(6) Winner's impression delivery
(7) "charge price" and "impression rendered" notification
User Tracking
(4) Retrieve user interests, geolocation, behavior
Data Hub:
DMP, Data Broker, etc.

Figure 1: High level overview of the RTB ecosystem. Several entities interact with each other, exchanging user's personal data before it is finally converted to money.

2 BACKGROUND ON RTB

RTB accounts for 74% of programmatically purchased advertising, reaching a total revenue of $8.7 billion in US [8], allowing advertisers to evaluate the collected data of a given user at real-time and bid for an ad-slot in the user's display. Next, we briefly cover the most important aspects of RTB auctions, key entities involved (§ 2.1), and how they are relevant to our study (§ 2.2).

2.1 The key players

As it can be seen in Figure 1, the key roles of the RTB ecosystem include the *Advertiser*, *Publisher*, *DSP*, *Ad-exchange*, and *SSP*, which interact with each other in several ways [4]. Note that it is very common for some (large) companies to play simultaneously different roles even inside the same auction (e.g. Google's DoubleClick Bid Manager [27] and DoubleClick for Publishers [23].

Publisher: (e.g., CNN.com) is the owner of a website, where users browse for content and receive ads (step 1). Each time a user visits the website, an auction takes place for each available ad slot. The ad impression of the winning advertiser is finally displayed in each auctioned slot of the website.

Advertiser: is the buyer of a website's ad slots. The advertiser creates ad campaigns and defines the audience that has to be targeted according to his marketing objectives, budgets, strategies, etc. In each auction, the one with the highest bid wins the ad slot and places its impression on the screen of the website's visitor.

Supply-Side Platform (SSP): is an agency platform, which enables publishers to manage their inventory of available ad slots and their pricing, allocate ad impressions from different channels (e.g. RTB or backfill in case of unsold inventory [46]) and receive revenue[1]. SSP is also responsible for interfacing the publisher's side to multiple ad-exchanges (step 2) and aggregate/manage publisher's connections with multiple ad networks and buyers. In addition, by using web beacons and cookie synchronization, SSPs perform user tracking in order to better configure their ad slots' pricing and achieve as many re-targeting ads as possible and thus higher bids [43]. Popular vendors selling SSP technology are OpenX, PubMatic, Rubicon Project, Right Media.

[1] Publishers can also interface directly with ADXs and handle their inventory on their own.

Winning Price Notification URLs
(A) cpp.imp.mpx.mopub.com/imp?ad_domain=amazon.es&ads_creative_id=ID&**bid_price=0.99**&bidder_id=ID&...&bidder_name=..&**charge_price=0.95**&country=..&...&currency=USD&latency=0.116&mopub_id=ID&pub_name=..
(B) tags.mathtag.com/notify/js?exch=ruc&...&**price=B6A3F3C19F50C7FD**&...&3pck=http%3A%2F%2Fbeacon-eu2.rubiconproject.com%2Fbeacon%2Ft%2Fce48666c-6eb4-46db-b0e9-6f4155eb557d%2F
(C) adserver-ir-p.mythings.com/ads/admainrtb.aspx?googid=ID&..&width=300&height=250&...&cmpid=ID&gid=ID&mcpm=60&...**rtbwinprice=VLwbi4K21KFAAAm2ziqnOS_O5oNkFuuJw**&..

Table 1: Examples of (A) cleartext, (B, C) encrypted RTB price notifications. "ID" is typically a hexadecimal number.

Ad-exchange (ADX): is a digital, real-time marketplace that, similarly to a stock exchange, enables advertisers and publishers to buy and sell advertising space through RTB-based auctions. ADX is responsible for hosting an RTB-based auction and distribute the ad requests along with user information it owns (i.e. browsing history, demographics, location, cookie-related info) among all the interested auction participants (step 3).

Typically these auctions follow the second higher price model (i.e. Vickrey auctions) [79], thus, the charge price for the winner of the slot is the second highest submitted. After the auction, the winning impression is served to the user's display within 100 ms of the initiating call (step 6) and the winning bidder is notified about the final charge price. Popular ad-exchanges include: DoubleClick, MoPub, and OpenX.

Demand-Side Platform (DSP): is an agency platform, which employs decision engines with sophisticated audience targeting and optimization algorithms aiming to help advertisers buy the best-matched ad slots from ADXs in a simple, convenient and unified way. DSPs retrieve and process user data from several sources (step 4) such as ADXs, Data Hubs, etc. The result of this processing is translated to a decision in practice: *How much is it worth to bid for an ad slot for this user, if any?* If the visitor's profile matches the audience the advertiser has focused his ad-campaign on, the DSP will submit to the ADX the impression and a bid in CPM (cost per 1000 impressions [38], typically in USD or Euro) on behalf of the advertiser (step 5). Popular DSPs are MediaMath, Criteo, DoubleClick, AppNexus and Invite Media.

User Data Hub, Data Exchange Platform (DXP): is a centralized data warehouse such as a Data Management Platform (DMP) [9, 45] or a Data Broker [49] which aggregates, cleans, analyzes and maintains user private data such as demographics, device fingerprints, interests, online and offline contextual and behavioral information [40, 41]. These user data are typically aggregated in two formats: 1) a full, audience user profile for offline analytics and data mining, 2) a run-time user profile, optimized for real-time requests such as RTB queries from DSPs, before submitting their bids to ADXs [18, 42] (step 4).

Such user profiles are sold to ad entities [5] because they increase the value of an RTB inventory by enabling a more behavioral-targeted advertising (2.7× more effective than non-targeted advertising [6, 82]). In fact, Data Hubs are considered the core component of the digital ad-ecosystem as they perform the attribution and labeling of users' data and create groups, namely *audience segments*, which are useful (i) to the publishers for their customer understanding, (ii) to the SSPs for retrieving more re-targeted ads and (iii) to the DSPs for feeding their bid decision engines. Further, quality scores are impartially assigned to users' private data based on the success of ad-campaigns they were used, thus driving the bid prices of future ad-campaigns. Notable DXPs are Turn, Adobe, Krux, Bluekai, Lotame.

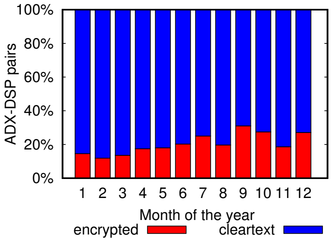


Figure 2: Portion of encrypted and cleartext pairs of ADX-DSP over time (2015).

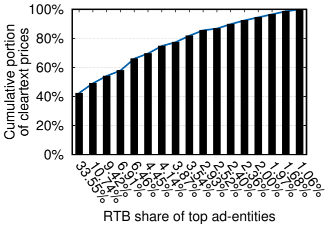


Figure 3: Cumulative portion of cleartext prices vs. ad-entities' portion of RTB.

2.2 RTB price notification channel

When an ADX selects the winning bid of an auction, the corresponding bidder must be notified about its win to log the successful entry and the price to be paid to the ADX. One could implement this notification in two ways: (i) with a server-to-server message between ADX and DSP, (ii) with a notification message conjoined with the price, passed through the user's browser as a call-back to the DSP.

The first option is straightforward and tamper-proof; no one can modify or block these messages, allowing companies to ensure that their logs are fully synced at any time. In addition, DSPs can hide information about the transactions, the purchased ad-slots and the prices paid from the prying eyes of competitors. However, DSPs do not have any indication of the delivery of each ad, in order to inform their campaigns and budget.

Instead, the second option not only can ensure the DSP that the winning impression was indeed delivered (the callback is fired soon after the impression is rendered on the user's device), but also gives the opportunity to drop a cookie on the user's device. Therefore, the second option is the dominant one in the current market: the ADX piggybacks a notification URL (nURL) in the ad-response, which delivers to the user the winning impression and the ad (steps 6 and 7 in Figure 1). This nURL includes basically the winning DSP's domain, the charge price, the impression ID, the auction ID and other relevant logistics (see Table 1 for some examples). In this present work, we study such nURLs and the prices embedded in them, as well as how they associate with the users' browsing behavior and other personal information.

2.3 Encrypted vs. cleartext prices

Although in the early years of RTB, all charge prices in nURLs were in cleartext, we see that nowadays more and more companies deliver charge prices in encrypted form (see examples in Table 1). While cleartext prices captured at the user's browser can be easily tallied to estimate the total cleartext cost, the same does not apply

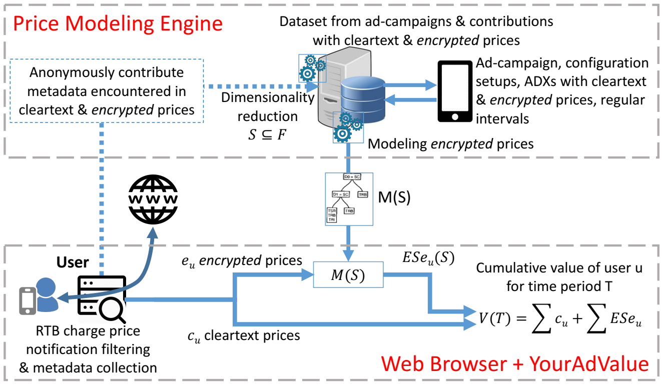

Price Modeling Engine
Dataset from ad-campaigns & contributions with cleartext & *encrypted* prices
Anonymously contribute metadata encountered in cleartext & *encrypted* prices
Dimensionality reduction $S \subseteq F$
Ad-campaign, configuration setups, ADXs with cleartext & *encrypted* prices, regular intervals
Modeling *encrypted* prices
M(S)
User
e_u *encrypted* prices
$M(S)$
$ESe_u(S)$
Cumulative value of user u for time period T
$V(T) = \sum c_u + \sum ESe_u$
RTB charge price notification filtering & metadata collection
c_u cleartext prices
Web Browser + YourAdValue

Figure 4: High level overview of our method. The user deploys YourAdValue on her device, which calculates in real-time the total cost paid for her by advertisers. In case of encrypted prices, it applies a decision tree model derived from the PME.

for the encrypted prices. The popular 28-byte encryption scheme companies use cannot be easily broken [26].

Previous studies [62] assumed that encrypted prices follow the same distribution as cleartext ones. Indeed, one may argue that the price encryption is just to avoid tampering of reported prices, so encrypted prices probably follow the cleartext price distribution. However, encryption provides also confidentiality to the bidding strategy. Thus, possible use of encryption in charge prices may be also a sign of a higher value that the bidder wants to hide: aggressive re-targeting because of user's previous incomplete purchases, targeting users with higher spending habits, or users with specialized needs (e.g., sensitive products, expensive drugs, etc.). Hence, a bidder (e.g. a DSP) may choose encryption to reduce transparency over its bidding strategies, or possible special knowledge it may have about a specific user, thus preventing an external observer or competitor from assessing its bidding methods and ad-campaigns.

We should note that encryption is not a feature that comes for free. There are significant costs for the participating parties such as more computation and storage overhead, energy consumption and higher imposed latency. Therefore, these costs alone could be a reason for an ADX to charge more for providing the benefits of encryption to a DSP. Considering all the aforementioned, in our study, we remove the need for making any assumptions regarding encrypted prices and allows us to account for any potential differences between cleartext and encrypted prices.

2.4 Encrypted prices on the rise

Encryption is a regular practice in desktop RTB ads (~68% as reported in [61] with major supporters being DoubleClick, RubiconProject and OpenX). By analyzing a weblog of 1600 real mobile users (see Section 4), we detected a smaller portion in mobile RTB ads (~26%). However, we found that the percentage of ADX-DSP pairs using encrypted price nURLs was steadily increasing through time (Figure 2), which means that more and more mobile advertising entities have started using nURLs with encrypted prices.

In fact, we found that the mobile advertising entities with the larger RTB shares deliver the highest portion of cleartext prices as well (Figure 3). For example, MoPub and Adnxs, the two leading ad-entities in our dataset, are responsible for 33.55% and 10.74% of the overall RTB ads detected, respectively (x-axis). They are also responsible for 45.40% and 5.45% for the cleartext prices detected, respectively (cumulatively in y-axis). If these two (or more) companies flipped their strategy from cleartext to encrypted, it would dramatically impact the RTB-ecosystem's transparency and hinder price information exposed to an external auditor or the involved user.

Notation	Definition
V_u	Total cost of user u
C_u, E_u	Sum of cleartext, encrypted prices of user u
SC_u, SE_u	Set of cleartext, encrypted price nURLs of user u
F_i	Vector of features for a price nURL i
$S_i \subseteq F_i$	Core features selected to represent nURL i
$ESe(S_i)$	Estimated encrypted price based on vector of features S of price nURL i

Table 2: Summary of notations.

Given these trends in mobile and desktop, we expect that in the near future RTB auctions will dominate, and many of the ad-entities will use encryption to deliver their charge prices. Our methodology anticipates these trends and promotes better transparency in online advertising and usage of user personal data. This methodology allows end-users to accurately estimate on their browser, at real-time, their average ad-related cumulative cost, even when the charge prices are encrypted.

3 METHODOLOGY

In this section, we describe our proposed methodology, with which a user u can estimate in real-time the accumulated cost V_u for the ads she was delivered while browsing the web (§ 3.1) (notations used are summarized in Table 2). Following this methodology, we design our system based on two main components: (i) a remote *Price Modeling Engine* (§ 3.2) and (ii) a user-side tool, namely *YourAdValue* (§ 3.3). Figure 4 presents an overview of our proposed methodology.

3.1 Overall cost of the user's data

The overall ad-cost of the user for time period T is the sum of charge prices the advertisers have paid after evaluating her personal data they have collected and delivering ads to her device. Specifically, this overall value is the sum of both her cleartext $C_u(T)$ and encrypted $E_u(T)$ prices and can be stated as:

$$V_u(T) = C_u(T) + E_u(T) \tag{1}$$

The cleartext prices of a user can be aggregated in a straightforward fashion, thus producing the ad-cost for user u over such prices:

$$C_u(T) = \sum_i c_i, \ \ where \ \ i \in SC_u(T) \tag{2}$$

On the other hand, the calculation of the aggregated $E_u(T)$ of the encrypted prices for the same user is not easy. The actual price values e_i are hidden and therefore need to be estimated. To achieve that, we leverage the metadata of each charge price in the user's set SE_u(T) of encrypted price notifications. Such metadata may include: time of day, day of week, size of ad, DSP/ADX involved, location, type of device, associated IAB, type of OS, user's interests, etc. All these metadata are collected in a feature vector F_i that captures the context of a specific charge price e_i in nURLi.

Metric	D	A1	A2
Time period	12 months	13 days	8 days
Impressions	78560	632667	318964
RTB publishers	∼5.6k/month	∼0.2k	∼0.3k
IAB categories	18	16	7
Users	1594	-	-

Table 3: Summary of dataset and ad-campaigns.

In order to estimate each encrypted notification price i, we built a machine learning model, which receives as input the feature vector F_i (or a subset $S_i \subseteq F_i$), extracted from the nURLi, and estimates a charge price $ESe(S_i)$ for the encrypted price e_i. This permits us then to aggregate the estimated encrypted prices for user u as we have done for the cleartext ones:

$$E_u(T) = \sum_i ESe(S_i), \ \ where \ \ i \in SE_u(T) \qquad (3)$$

3.2 Price Modeling Engine

The core element of our solution, the Price Modeling Engine (PME), is a centralized repository responsible for the estimation of encrypted prices. To achieve this, the PME requires a sample of charge price data and associated features to train a machine learning model. This component is designed to incorporate data such as offline weblogs (see Section 4), or online anonymous contributions (anonymized features and charge prices) from participating users, similarly to other systems that depend on crowd-sourcing (e.g., Floodwatch [76]). Using such data, the PME can re-train the computed model at any time. To assess the difference between cleartext and encrypted price distributions in the wild and fine-tune the training model, the PME runs small "probing ad-campaigns" to collect ground truth of real charge prices from both encrypted and cleartext formats.

Feeding the PME with all possible metadata available, i.e. auctions' metadata and users' personal data, is clearly not practical. There exist hundreds of data points per individual price. Passing all of them to the modeling engine would make the computational cost excessive. Additionally, if all data points were to be used in the probing ad-campaigns, they would render such campaigns too expensive for their purpose. In order to run effective and efficient ad-campaigns, and allow the training of a price model without high computation overhead, the PME performs careful dimensionality reduction on the extracted metadata (F) to derive a subset $S \subseteq F$ of core features capable to capture the value of an impression. This dimensionality reduction makes the probing ad-campaigns feasible by reducing by many orders of magnitude the needed features of each testing setup, and effectively the number of setups to be tested (see Section 5).

Using the collected ground truth of encrypted prices from ad-campaigns, the PME trains a machine learning model M to infer encrypted prices based on their associated subset of features S. Then, each user can apply the model M (in the form of a decision tree) locally on their device to estimate each of her encrypted charge prices based on the matching metadata S.

In case the availability of cleartext prices is limited, the reduction step to identify important features could be hindered, but not obstructed. To mitigate this, the PME can run more probing ad-campaigns to cover extra features found in users' anonymous contributions, or that are available in professional ad-campaign planners (as in FDVT [14]). Then, the most important features can be selected based on their contribution to model the encrypted prices extracted from these campaigns.

Type	Feature
Geo-temporal	Time of day, Day of week
	Location of user based on IP, # of unique locations of the user, location history
User	Interest categories of the user, Type of mobile device, # of total web beacons detected for the user, # of cookie syncs detected of the user up to now, # of publishers visited by the user, # of total bytes consumed by the user,
	Avg. number of reqs per user for the advertiser, # of HTTP reqs of the user, Avg. number of bytes per req of user, Total duration of reqs of the user, Avg. duration per req of the user
Ad	Size of ad, ADX of nURL, DSP of nURL, IAB category of the publisher, popularity of particular ad-campaign,
	# of total HTTP reqs of the advertiser, # of bytes of HTTP req, Avg. duration of the reqs for the advertiser, # of URL parameters, Number of total bytes delivered for the advertiser

Table 4: Features extracted by summarizing data from parameters embedded in each price notification detected in the dataset for users and advertisers.

3.3 YourAdValue

YourAdValue is a user-side tool responsible for monitoring the user's nURLs and calculating locally the cumulative cost paid for her in real-time. To achieve this, it filters nURLs from her network traffic and extracts (i) the RTB auction's charge prices (both encrypted and cleartext), and (ii) metadata from each specific auction (e.g. time of day, day of week, size of ad, involved DSP and ADX, etc.) along with the personal data the user leaks while using online services (e.g. location, type of device and browser, type of OS, browsing history, etc.).

As we mentioned earlier, cleartext prices can be aggregated directly, but encrypted prices must be estimated. Therefore, YourAdValue retrieves from the PME a model $M(S_i)$ that (i) includes the features S_i that need to be extracted from the collected metadata, and (ii) provides a decision tree for the estimation of an encrypted price based on these features.

Using this model, YourAdValue can estimate locally on the client side, the value $ESe(S_i)$ of the encrypted charge prices based on the features S_i of the given nURL. After estimating each encrypted price, YourAdValue presents to the user the calculated sums $C_u(T)$ and $E_u(T)$ along with relevant statistics and the total amount $V_u(T)$ paid by advertisers (see Section 6).

YourAdValue can be implemented in the same manner, either as a browser extension for desktops or as a module for mobile devices. In the latter case, YourAdValue can monitor traffic of both browsers and apps similar to existing approaches [64]. For simplicity, in this work we design YourAdValue as a browser extension; its mobile counterpart is part of our future work.

Our tool, built as an extension for Chrome browser, monitors both HTTP and HTTPS traffic of the user and detects the RTB nURLs. Additionally, it stores in the browser's local storage the

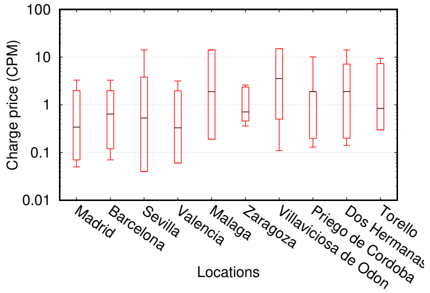


Figure 5: Distribution of charge prices per city (sorted by city size).

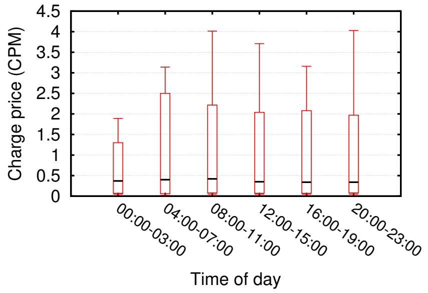


Figure 6: Distribution of charge prices for different times of day.

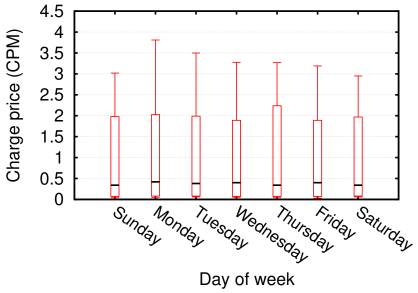


Figure 7: Distribution of charge prices for different days of week.

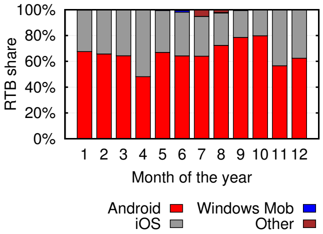


Figure 8: Portion of RTB traffic for top mobile OSes.

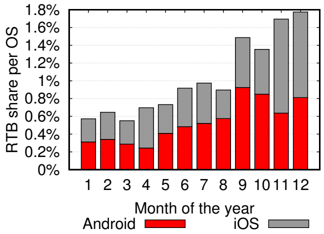


Figure 9: Portion of RTB traffic normalized by OS.

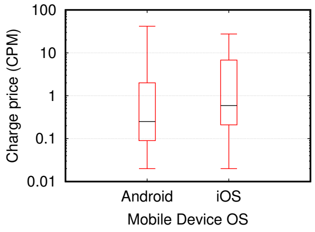


Figure 10: Distribution of charge prices per mobile OS.

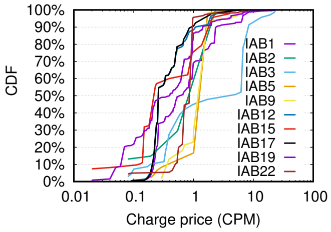


Figure 11: CDF of the generated cost per IAB category.

filtered charge prices, the personal and auctions metadata and the estimation of the encrypted prices. The extension, through toolbar notifications, informs the user about newly detected RTB charge prices. Upon request, it reports the cumulative cost along with the previous individual charge prices. Finally, the extension periodically issues requests to PME to check for new versions of the model.

4 BOOTSTRAPPING PME

We assess the feasibility and effectiveness of our methodology by bootstrapping the PME to train our model on real data by collecting a year long dataset containing weblogs from 1594 volunteering mobile users from the same country. Our users agreed to use a server of our control as a proxy, allowing us to monitor their outgoing HTTP traffic.[2] As a result, we were able to collect a large dataset D of 373M HTTP requests spanning the entire year of 2015. Note that though our dataset consists of HTTP-only traffic, in principle our approach works with HTTPS as well, using as input the users' contributed data as can be seen in Figure 4. Table 3 presents a summary of our dataset D. Next, we present the data collection and analysis to extract features used in the price modeling and ad-campaign planning.

4.1 Dataset analysis

Weblog Ads Analyzer. To process our dataset, we implemented a weblog advertisements analyzer, capable of detecting and extracting RTB-related ad traffic. First, the analyzer uses a traffic classification module to categorize HTTP requests based on an integrated blacklist of the popular browser adblocker Disconnect [15].[3] Using this blacklist, the analyzer categorizes domains in 5 groups based on the content they deliver: (i) Advertising, (ii) Analytics, (iii) Social, (iv) 3rd party content, (v) Rest. It consequently applies a second-level filtering on the advertising traffic by parsing each URL for any RTB-related parameters (like nURL). The analyzer detects nURLs by applying pattern matching against a list of macros we collected after (i) manual inspection and past papers [53, 62], and (ii) studying the existing RTB APIs [25, 35, 56, 63, 69] used by the current dominant advertising companies. From these detected nURLs, it extracts the charge prices which we assume in this study that are in US dollars[4] paid by the winning bidders, after filtering out any bidding prices that may co-exist in each nURL. It also extracts additional ad-related parameters such as ad impression ID, bidder's name, ad campaign ID, auction's ad-slot size, carrier, etc.

Other operations carried out by our analyzer include: (i) user localization based on reverse IP geo-coding, (ii) separation of mobile web browser and application originated traffic based on the `user-agent` field of each HTTP request, (iii) extraction of device-related attributes from the `user-agent` field (type of device, screen size, OS etc.), (iv) identification of cooperating ADXs - DSPs pairs, leveraging the nURL used by the ADX to inform the bidder (i.e. DSP) about its auction win, (v) user interest profile based on web browsing history.

Feature extraction. DSPs use different machine learning algorithms for their decision engines, taking various features as input, each affecting differently the bidding price and, consequently, the charge price of an ad-slot. To identify such important parameters, we extracted several features from the nURLs of our dataset such as user mobility patterns, temporal features, user interests, device characteristics, ad-slot sizes, cookie synchronizations [1],

[2] Data were treated anonymously although users signed a consent form allowing us to collect and analyze their data.

[3] Our analyzer can also integrate more than one blacklists (e.g., Adblock Plus' Easylist, Ghostery's blacklist, etc.)

[4] Given that the majority of ADXs are located in US and following previous works [62], we assume every charge price to be in US Dollars (so $1CPM$ =1/1000 impressions).

publisher ranking, etc. Next, we present the analysis of the most interesting features (Table 4 presents a summary). We group them into 3 categories: geo-temporal state of the auction (§ 4.2), user’s characteristics (§ 4.3), and ad-related (§ 4.4).

4.2 Geo-temporal features

An important parameter that affects the price of an RTB ad is the user’s current location [31], information which is broadly available to publishers and trackers. Thus, in our dataset we extract user IP address and using the publicly accessible MaxMind geoIP database [54], we map each IP to its city level. In Figure 5, which presents the 5th, 10th, 50th, 90th and 95th percentile of the charge prices, we see that although the median values are relatively lower in large cities, the fluctuation of their price values is higher.

Another important feature is time, and specifically the time of day and day of week. This is important due to the different level of attention a user may give to an ad impression and the amount of time she has to purchase an advertised product (e.g., working hours vs. afternoon’s free time, or weekdays vs. weekends). In Figure 6, although the median charge prices are of similar range, we see that the early morning hours until noon tend to have more charge prices with increased values. In Figure 7, we see a periodic phenomenon, where although in median values the charge prices are quite close, during weekdays the max prices are relatively higher than on weekends.[5]

4.3 User-related features

Device type. By parsing the user-agent (UA) header information, our analyzer classifies traffic and inspects the different fingerprints the UA leaks (specifications of process virtual machine (e.g., Dalvik or ART) or kernel (e.g., Darwin), operating system, browser vendor etc.) Thus, we are able to identify the type of device (PC or mobile), the different types of mobile operating systems (Android, iOS, Windows) and if the traffic was generated from a mobile app or a mobile web browser.

In Figure 8, we see the percentage of RTB traffic for the different OSes over time. As expected, Android and iOS dominate, owning the larger portions of the market through the entire year, with Android-based devices appearing in 2x times more RTB auctions. However, when normalizing this RTB share per mobile OS (Figure 9), we find that Android and iOS devices are delivered mostly equal RTB impressions, with some months Android surpassing iOS and vice-versa. Then, we extract the traffic originated from the most popular ad-entity, MoPub [57], and analyze the charge prices of the impressions rendered in the different OSes. Surprisingly, although Android-based devices are more popular, we see in Figure 10 that iOS-based devices tend to receive higher RTB prices, in median values.

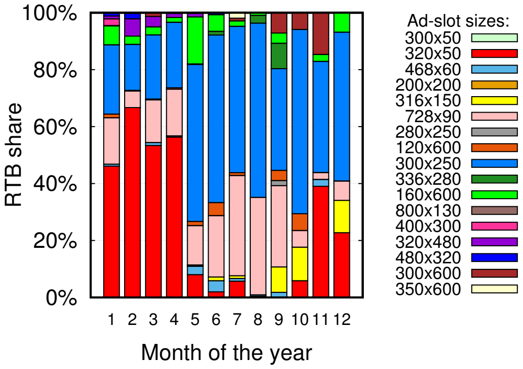


Figure 12: Ad-slot size popularity through time (sorted by area size).

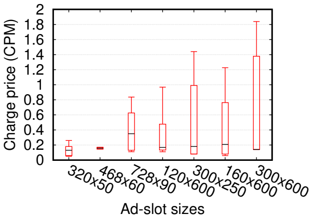


Figure 13: Distribution of the charge prices per ad-slot size (sorted by area size).

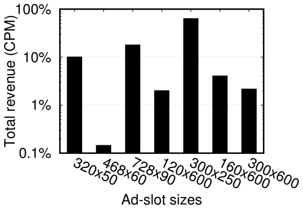


Figure 14: Accumulated revenue per ad-slot size (sorted by area size).

Inference of the user’s interest. The browsing history of a user is used by the advertising ecosystem as a proxy of her interests. By monitoring the websites a user visits through time, a tracker can infer her interests, political or sexual preferences, hobbies, etc., quite accurately [7]. To enrich our set of features with the users’ interests, we collect all the websites each user visits across her whole network activity. Such information is available to the RTB ecosystem as well, by using cookie synchronization [1] or web beacons [34]. To extract the interests from the visited websites, similar to existing approaches [3], we retrieve the associated categories of content for each website according to Google AdWords [28]. Then, we aggregate across groups of categories for each user and get the final weighted group of interests for each user in the form of IAB categories [37]. Figure 11 presents for the top mobile ADX (MoPub) a distribution of the generated ad revenue for the different IAB content categories in a 2 month subset of our dataset. As expected, not all IAB categories cost the same. Indeed, there are categories that are associated with products which attract higher bid prices in auctions, like IAB-3 (Business & Marketing), with an average charge price of up to 5 CPM for the 50% of the cases. Alternatively, there are categories like IAB-15 (Science), which are unable to draw prices higher than 0.2 CPM for the 50% of the cases.

4.4 Ad-related features

Web Vs. Apps Advertisers bid for ad-space in both webpages and mobile apps. After studying the cost per ad in both counterparts in our dataset, we see that apps draw on average 2.6× higher prices (0.712 CPM vs. 0.273 CPM). This is expected; studies have shown that more advertising budget is spent on mobile application ads instead of mobile web, driving higher prices per ad [58]: (i) Users pay more attention to app ads as they typically occupy fixed places in the screen, with no opportunity to scroll them out of sight as in web ads. In addition, (ii) studies [64] have shown that apps leak more personal data to advertisers, enabling them to deliver more targeted ads.

[5] For time-of-day and day-of-week distributions, which visually appear to be similar, we confirmed that they are, in fact, statistically different with non-parametric, two-sample Kolmogorov-Smirnoff tests at p-value levels of $p_{tod}<0.0002$ and $p_{dow}<0.002$.

Ad-slot sizes. Some ad-entities carry in their nURLs a parameter with the size of the auctioned ad-slot. In Figure 12 we plot the popularity of each of the ad-slot sizes through time. It's interesting to see that 300x250 ad-slots (known as "MPUs" or "Medium Rectangles") dominate the dataset from May'15 on, taking the place of 320x50 ad-slots (known as "large mobile banners"). In fact, 300x250 ad-slots have more ad content available from advertisers, so they can increase earnings when both text and image ads are enabled. In addition, we see that the 728x90 ad-slot ("leaderboard" or "banner") is also popular. This ad-slot, usually placed at the top of a page, is seen by users immediately upon page load.

It is easy to anticipate, that the more space an ad-slot covers in the user's display, the higher the price will be. To verify this intuition, we isolated the traffic of an ad-entity (i.e. Turn [18]), which carries the ad-slot size in its nURLs along with the associated charge prices. Surprisingly, in Figure 13, we see that this intuition is wrong since the most expensive ad-slots for an advertiser are in fact, not the largest ones. In our dataset, we see that the two most expensive ad-slots are the MPU (300x250) and Monster MPU (300x600), with median prices of 0.47 and 0.39 CPM, respectively. However, from Figure 14, the increased popularity of MPU and Leaderboard ad-slots, allows them to accumulate 64.3% and 20.6% of the total RTB revenue of Turn in our dataset, respectively. Finally, it is worth to note that our results verify past resources [16, 24] regarding the more expensive ad-slots.

4.5 Summary

In summary, by analyzing the features extracted from our offline dataset, we find that a user's location (at city level) affects the median price that advertisers pay as well as its variability. However, such price differences are expected to be more evident at the country-level, as shown in [62]. In addition, the days and hours that a user may not be busy (Sundays), or may offer more attention (e.g., early mornings, Mondays) lead to higher charge prices. The type of user's device also affects the charge prices but in a rather contradicting fashion: though there are more Android devices, iOS-based devices draw higher median prices. As expected, the total revenue per category of user interest (through IABs) varies a lot, with some IABs being more costly than others. Finally, the display's real-estate occupied by an ad-slot does not correlate well with price. In fact, larger ad-slot sizes do not mean higher prices. As shown in the next section, these extracted features are used to plan effective ad-campaigns and model encrypted charge prices.

5 CHARGE PRICE ESTIMATION

In order to create a model that estimates the encrypted prices detected on the user's browser and computes the total cost advertisers pay for her personal data, we need to have ground truth on charge encrypted prices. However, such dataset is not easy to acquire. One way to obtain this information is to collaborate directly with an ADX that sends such encrypted prices.[6] We assume this to be the rare case, since ADXs are generally unwilling to share such kind of data that may reveal bidding strategies and revenues.

In order to collect ground-truth data on encrypted prices, our system conducts small probing ad-campaigns on ADX(s)-DSP pairs that encrypt the winning prices. Such ad-campaigns can be designed and executed with the help of a single or few DSPs, with little overhead and a small budget of a few hundred dollars. In addition, they can be optimized by using a specific set of experimental setups, which cover all possible scenarios from the small parameter vector S to be kept short, efficient and cheap. Given that the prices do not change drastically over time, these campaigns can be executed every few months to collect probing data for *time-shift correction* and increased coverage of more ADXs. Besides, they can be automated and re-launched as frequently as needed, e.g., every few months or when the detected cleartext prices deviate from historical data. Having such campaigns launched from a specific location allows for more accurate and cost efficient price modeling that can be shared across all participating platform users in the same area or country.

We envision that such campaigns can be crowd-funded (like Tor Project [77], Wikipedia, WiGLE [10], etc.), thus, contributing to an independent and sustainable platform that can scale better across users, countries, and ADXs covered. One may argue that these probing campaigns could pollute users' browsing with non-useful ad impressions. Thus, they need to comply with the current standards, and if possible, consider an actual product or service. Of course, ADXs could in principle fight back and try to identify and block such campaigns, but their huge clientele combined with the low volume of such campaigns makes the detection very difficult. Next, we describe the effort to select a subset of core features important for price modeling (§ 5.1) and how they allow us to design efficient and effective ad-campaigns (§ 5.2). Then, we provide an analysis of the data collected by two such campaigns (§ 5.3) and we describe the model that estimates encrypted prices, which can be used by end-users (§ 5.4).

5.1 Dimensionality reduction of features

The cost of testing all possible combinations of parameters and their values from the available feature set F (with one probing ad-campaign each), would constitute the budget for the ad-campaigns impossible (1000s of setups x 10s Euros/setup). Therefore, to perform ad-campaigns that are both effective and cost efficient, we need to select a subset of features $S \subseteq F$ that best describe the RTB prices found in weblogs such as the historic dataset D. This subset of features should explain as much of the variability of prices as possible. Assuming both encrypted and cleartext prices are affected by the same set of important features, this set should be small. The fewer features we select as important, the smaller the cost of running ad-campaigns to collect representative RTB prices using these features (e.g., 10s-100s of setups).

To achieve this selection, we performed dimensionality reduction using all the available features (288) described in Section 4 and Table 4, using the cleartext prices as the target variable for optimization. Some of these features are dense, i.e., they have an actual value in each price (e.g., time of day, day of week, size of ad, etc.) and others are sparse (e.g., interest categories of the user through time, publishers visited by user through time, etc.). First, for normalization, we applied a log transformation on the extracted

[6] We considered top ADXs for encrypted prices (DoubleClick, OpenX, RubiconProject, PulsePoint), and ADXs for cleartext prices like MoPub (top mobile ADX).

Filter name	Range of values (type)
Cities	Madrid, Barcelona, Valencia, Seville
Type of interaction	Mobile in-app, Mobile web
Time of day	12am-9am, 9am-6pm, 6pm-12am
Day of week	Weekday, Weekend
Type of device	Smartphone, Tablet
Type of OS	iOS, Android
Ad-format (smartphone)	320x50, 300x250, 320x480 or 480x320
Ad-format (tablet)	728x90, 300x250, 768x1024 or 1024x768
Ad-exchange	MoPub, OpenX, Rubicon, DoubleClick, PulsePoint
Categories of targeting	all IABs possible

Table 5: Basic filters used in controlled ad-campaigns in Spain. In total, 144 experimental setups were attempted.

cleartext prices found in D. Then, we applied a clustering of the prices into 4 classes, using an unsupervised equidistance model that finds the optimal splits between given prices using a method of leave-one-out estimate of the entropy of values in each class. Next, we filtered out features that did not vary at all (i.e., constants) or had very high variance (99%) (i.e., likely to be noise).

As a final step, dimensionality reduction (or feature selection) techniques such as PCA or Random Forests (RF) can be used [44]. We chose the RF model[7] because it takes into account the target variable (cleartext price), it can be trained quickly on large datasets, it maintains interpretability of features and generally does not overfit the given data. In case the availability of cleartext prices is limited, the reduction step to identify important features to be used in ad-campaigns could be hindered. To mitigate this, the PME can use intermediate techniques such as high correlation filters that do not require a target variable, to eliminate features carrying similar information.

We trained various RF models using subsets of semantically related features from the available feature set and the best features from each subset were selected based on their power to describe the cleartext price distribution. In summary, we grouped features in the following sets: A) time, B) http-related, C) advertisement-related, D) DSP-related, E) publisher/host interests, F) user http statistics (historical), G) user interests (historical), and H) user locations (historical). We also tried selecting representative features out of each set to create minimal combinations covering all aspects of the http-available information.

In total, we tried tens of feature subsets and combinations and evaluated them using standard machine learning metrics such as precision, recall, weighted area under the receiver operating characteristic curve (AUCROC) and out-of-bag error. Dimensionality reduction could, in principle, lead to loss of accuracy in the effort to explain price classes. However, our experimentation lead to a small subset of features with minimal loss of precision (< 2%) and recall (< 6%). In fact, we conclude that an optimal subset that performs very well and is small enough to allow cost efficient ad-campaigns is a set that *combines features from different groups*. In particular, also confirmed with an ad-campaign expert, we select the following features to be used for the probing ad-campaigns described next:

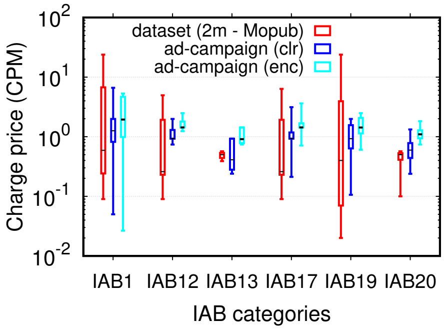


Figure 15: Comparison of CPM costs for the different IAB categories in our dataset and the 2 probing ad-campaigns.

S={application/web-browsing, device type, user location, time of day, day of week, ad format (size), type of website, ad-exchange}.

5.2 Ad-campaigns setup

Using the most important parameters extracted in set S, we construct various experimental setups $s \in S \subseteq F$ that can be used to deploy such ad-campaigns over a short period of time T' in selected ADXs to match top ADXs found in D. These setups combine different values of control variables that are important for an ad-campaign:<user location, web-interaction type, time of day, day of week, device type, OS, ad-size, ADX>. For example, an experimental setup could be this: <Madrid, app, 12am-9am, weekday, smartphone, iOS, 320x50, MoPub> (144 setups, Table 5). Clearly, using more features would increase coverage of different types of ads, but also the campaigns' cost. Instead, by running controlled ad-campaigns with a small feature set, we can receive ground truth data about encrypted prices, thereby allowing us to train a model for such prices, in a reasonable ad-campaign cost.

Campaigns with ADXs that deliver cleartext prices also allow us to compare prices in different times and compute shifts in the price distribution due to time passed between the collection of dataset D and present time. To compensate for the loss of information from cleartext prices becoming less abundant, additional features available in professional ad-campaign planners (as in FDVT [14]) could be used in the future to enhance the setups tested. With the results of these campaigns (in essence, charge prices for RTB ads that fulfil a given setup s), the PME can train a model to estimate the cost of new ads with a given setup s' close, or equal, to one tested, i.e. $s' \sim s \in S$.

Number of required ad-campaigns. An important decision in running probing campaigns is how many of them to launch, and with how many impressions in each one, in order to obtain a good approximation of the underlying distribution of prices. For this, we analyzed the ad-campaigns found for MoPub in D. We identified 280 such campaigns in 2015, with mean and standard deviation of charge price of $m = 1.84$ and $std = 2.15$ CPM, respectively. We use the process described in [39] and the next formulation to compute d, the expected error on the mean, assuming a suggested number of setups n, and ignoring the finite population correction adjustment (thus assuming a more conservative approximation of n) $d = \frac{Z_{\alpha/2} \times std}{\sqrt{n}}$, where Z is the z-score of normal distribution. Using the 144 setups proposed, we can approximate to more than 95% CI (i.e., α=0.05) the mean price of campaigns observed in the

[7] An ensemble of decision trees built using a random subset from the available features.

wild, assuming a margin of error 0.35 CPM. Also, considering the distribution of prices within the largest of ad-campaigns detected for MoPub with 1.8k impressions, we can approximate to 95% CI the mean price of a campaign, assuming an error 0.1 CPM and minimum of 185 impressions per campaign.

5.3 Ad-campaigns analysis

Using the above as guideline, we executed two different ad-campaigns to collect data on prices (Table 3). Our ad-campaigns advertised a real non-for-profit NGO in the area of data transparency, in an attempt to avoid polluting users with meaningless impressions, and trying to do something useful with the allocated budget.

Dataset collected. The first round ($A1$) was executed for 2 weeks in May 2016 and utilized the 4 ADXs mentioned earlier (also found in D) that encrypt price notifications and targeted publishers of many IAB categories. The second round ($A2$) was executed with the same experimental setups as $A1$ during June 2016, but in this case the DSP was instructed to use only MoPub, while still targeting similar IAB categories of publishers. These constraints allowed us to directly compare encrypted with cleartext prices in the same period, and time-shift all prices detected in D from 2015 to 2016.

In both campaigns, the DSP was given an upper bound on the bidding CPM price to safeguard that the allocated budget will not be consumed quickly. Because studying the effects of retargeting is beyond the scope of this paper, we did not ask the DSP to perform such campaigns. However, the DSP was instructed to bid in a dynamic manner, as low or high as needed to get the minimum of impressions delivered for the various experimental setups we requested. We plan to investigate the effects of retargeting in a separate and dedicated future study. Overall, we managed to receive across all setups, over $600k$ impressions displayed with encrypted price notifications to more than 200 publishers, and over $300k$ impressions with cleartext price notifications to more than 300 publishers, reaching audiences of 6 IAB categories common to both notification types.

Cost paid vs. IAB category. In Figure 15, we compare the overlapping IAB categories of the RTB impressions we took from (i) the set of encrypted prices from the ad-campaign on four ADXs in $A1$, (ii) the set of cleartext prices from the ad-campaign on MoPub ($A2$), (iii) the 2 months MoPub subset of D. Note that in some cases, the results from D vary more than in the ad-campaigns. This is to be expected, as the dataset includes prices from numerous DSP-ADX pairs for many ad-campaigns running in parallel in the duration of a year, whereas our two ad-campaigns are more targeted to specific DSP-ADX pairs.

Regarding the cleartext prices of different IAB categories, although the median prices are usually in the same order of magnitude, they are higher in the case of the recent ad-campaign contrary to the 2 month dataset. We believe that this difference is due to the time shift between the dataset collected in 2015 and the ad campaigns performed in 2016. In addition, we see that the median price is always higher in case of encrypted prices ($A1$), compared to the cleartext prices of the second ad-campaign ($A2$) and dataset D.

5.4 Encrypted price modeling

Using the ground truth data collected from the first round of ad-campaigns (encrypted prices) with various parameters within the subset of features S, we trained a machine learning classifier to predict values of encrypted prices. We note that given the problem of modeling real values, we first applied regression models with different combinations of dependent variables (S). However, the high variability of charge prices lead to low performance (high error) of the regression models. Therefore, we proceeded to split the prices into groups for classification. As a first step, we performed similar preprocessing for the encrypted prices as we did earlier for the cleartext prices (normalization and clustering to 4 classes of well balanced groups). Next, we trained a RF model to predict the class of an encrypted price, based on the available parameters S. For the training and testing, we applied 10-fold cross validation, and averaged results over 10 runs. Using features such as city of user, day of week and the time the ad was delivered, ad size, mobile OS of the user's device, IAB category of the publisher, ADX used and device type, our classifier can achieve a very good performance: TP=82.9%, FP=6.8%, Precision=83.5%, Recall=82.9%, 0.964 AUCROC. These scores are weighted averages across all classes, with no class performing worse than 5% from the average. We repeated this process with more price classes (i.e., 5-10 groups) for higher granularity of price prediction, but the results with 4 classes outperformed them.

When the exact publisher used is also taken into account in the model, the performance of the classifier increases to 95% accuracy, and 0.99 AUCROC. However, this is classic over-fitting and we should caution that the publishers used in the ad-campaigns are just a subset of the thousands of possible publishers that can be found in real weblogs. Therefore, we chose to use the model with the IAB category but without the exact publisher as part of its input features. Next, this model was used for the estimation of the encrypted prices of nURLs found in the weblogs of each user in D, given the matching parameter values from $S \subseteq F$.

6 USER COST FOR ADVERTISERS

The previous sections allowed us to: (1) bootstrap our price modeling engine from existing user weblogs, so that we find the important features describing well the observed RTB cleartext prices, (2) using these important features, run probing ad-campaigns with ADXs that send encrypted price notifications, so that we collect ground truth on such prices from performance reports delivered to us, (3) using such ground truth, train a machine learning model to estimate the price of new RTB notifications sent in encrypted form. We are now ready to study the overall cost advertisers paid for each of the users in our dataset D, who received cleartext and/or encrypted prices in nURLs of delivered ads.

6.1 Encrypted vs. cleartext price distributions

The work in [62] assumed that encrypted prices follow the same distribution with cleartext prices. To examine the validity of such assumption, we plot the distributions of both encrypted and cleartext charge prices we got from the two ad-campaigns we performed. Interestingly, from Figure 16, the distribution of *encrypted prices* in $A1$ is distinctly different and of *higher median value (~1.7×) than cleartext prices* of $A2$.

In addition, we study the distributions between different time periods and ADXs to extract important lessons. First, we see that

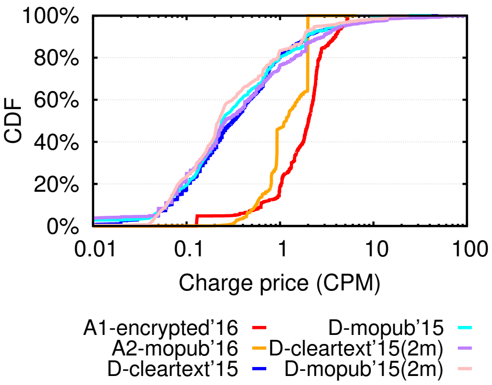


Figure 16: Comparison of price distributions between cleartext and encrypted, for different time periods and datasets (D vs. $A1$ and $A2$).

the cleartext price distribution of MoPub (2015) is similar to all ADXs sending cleartext prices, either when considering a 2 month period or a full year. Hence, we can study MoPub as a representative example and extrapolate lessons for the rest of the ADXs that send cleartext prices. Second, the distribution of cleartext prices from $A2$ (MoPub) are of higher median value and can be used to establish the price shift due to time difference between the time T the dataset was collected, and T' when the campaigns were executed. In reality, this price shift can be detected evenly across multiple probing ad-campaigns (e.g., once per quarter of year).

6.2 How much do advertisers pay to reach a user?

Equipped with our presented methodology for estimating encrypted prices, we are now ready to respond to our motivating question. Specifically, we utilize our method and compute the overall cost advertisers paid for each user in the dataset D, i.e., across a whole year of mobile web transactions. We also apply a time-correction coefficient on the cleartext prices using the prices from the second round of ad-campaigns. This allows us to consider the increase in cleartext prices due to time difference from the weblog collection (2015) and the ad-campaigns execution (2016).

Figure 17 presents these cumulative costs in the form of CDFs of the price distributions. As expected, we observe that the cumulative cost due to encrypted prices is still not surpassing the cleartext, since the latter is still the dominant price delivery mechanism in mobile RTB. We also note that some users are more costly than others. Specifically, the median user costs ~25 CPM, and up to 73% of the users cost $<$ 100 CPM through the year for the mobile ad ecosystem in the given dataset. This means that the ad-ecosystem reaches the average user very cheaply and multiple times below what users estimate this cost to be (e.g.,10s of dollars [11]).

On the other hand, for ~ 2% of users, the advertising ecosystem spent 1000-10000 CPM for the same time period. Finally, about 60% of users had an increased average cumulative cost of ~ 55% on top of their cleartext cost, due to the estimated encrypted prices. These users had a median of 14.3 CPM added to their total cost, with some extreme cases of 1000-5000 CPM.

In the previous result, we compared the distributions of encrypted and cleartext prices, while disregarding the targeted user.

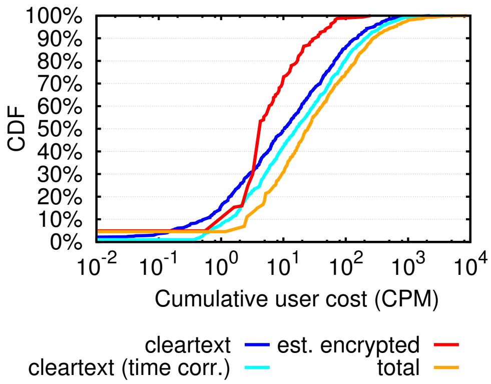


Figure 17: Cumulative CPM paid per user in our year long dataset.

In order to identify if the cost paid through encrypted prices is the same with cleartext for a specific user, we compare for each user the total costs in Figure 18 and average cost per impression in Figure 19, for each type of price. We observe that a significant portion of users (~20-25%) cost similarly for ads embedded with encrypted or cleartext prices. As expected, due to the current majority of cleartext prices in the mobile ad market, a large portion of users (~75%) have higher cumulative cost from cleartext than encrypted prices. However, a small portion (~2%) costs more (2-32×) in encrypted than in cleartext form, because they were delivered mostly ads with encrypted prices. When we normalize the cumulative ad cost of user per impression delivered (Figure 19), we find that for small prices of ≤3 CPM/impression, cleartext is more dominant across users. We also find a small portion (~2%) of users who cost up to 5× more CPM/impression for the delivered ads in encrypted than in cleartext form. We anticipate this portion to increase as the encrypted notification becomes the dominant delivery of RTB prices in mobile.

6.3 Summary

By studying the overall RTB advertising cost for users in our dataset, and distinguishing the encrypted from the cleartext prices, we found that the basic assumption of related work [62] that encrypted and cleartext prices are similar, is not valid (encrypted prices are around 1.7× higher). Furthermore, advertisers, based on users' personal data, paid ~25 CPM for delivering ads to an average user, and less than ~100 CPM for delivering ads to 3/4 of users during a year. We also identified a small portion of outlier users (~2%) who cost 10-100× more to the ad-ecosystem than the average user, and a similar portion that costs up to 32× more in encrypted than cleartext prices, even though encrypted prices are only a quarter of the mobile RTB ecosystem.

Validation. As an effort to validate our methodology, we can extrapolate how much users cost for the ad-ecosystem and if this estimation compares with current market numbers. For this extrapolation, we make some assumptions on how our dataset represents the overall ecosystem of users and advertisers. In particular, we assume that our average mobile user, whose annual ad-cost is in the 8-102 CPM range (25th-75th perc.), has: (1) performed 2.65 hours online daily, which is ~83% of the average daily mobile internet usage, when considering average tablet and other mobile device

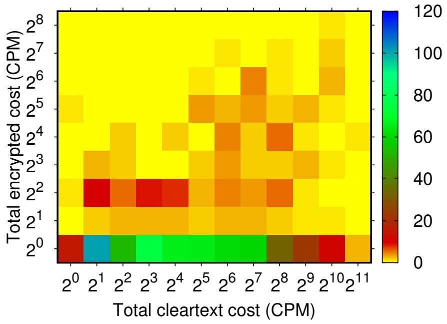


Figure 18: Total cleartext vs. total estimated encrypted cost of each user in D (color indicates number of users).

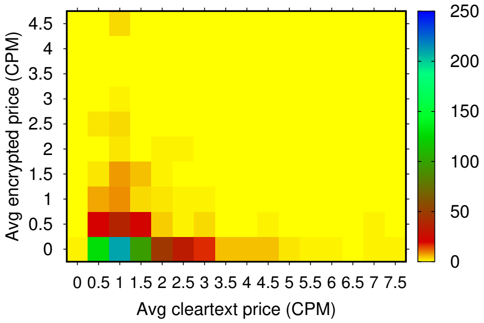


Figure 19: Average cleartext vs. average estimated encrypted price per impression of each user in D.

usage [50], (2) performed internet activity from both mobile and laptop/desktop devices, the former traffic type being ~51% of total internet time [12], (3) received ads in a similar fashion in both HTTP and HTTPS, the former being ~40% of the total traffic delivered to a user [20, 72], (4) received ads over RTB, which has an overhead management and intermediaries cost of ~55% [68], and (5) received ads in a similar fashion over RTB and traditional and other online advertising, the former being ~20% of the total online advertising [36]. Considering these factors, the overall average user ad-cost (25th-75th perc.) would be in the range of \$0.54-6.85, which is in the order of magnitude reported by major online advertising platforms such as Twitter (owner of MoPub, ARPU: \$7-8 [30]) and Facebook (ARPU: \$14-17 [13]) during the period 2015-2016.

7 RELATED WORK

There is a plethora of papers studying privacy loss and tracking techniques in the wild [1, 17, 21, 48, 52, 60, 71]. There are also others proposing privacy preserving countermeasures based on either (i) randomization-/obfuscation- based techniques [59, 65], where the authors aim to pollute the information trackers retrieve in order to hide the users' data and interests, or (ii) anti-tracking mechanisms [47, 64], where requests to trackers are avoided or blocked. All the above studies, highlight the voracity of web entities to collect data about the user and her online behavior, and an arms race between the privacy-aware users and trackers.

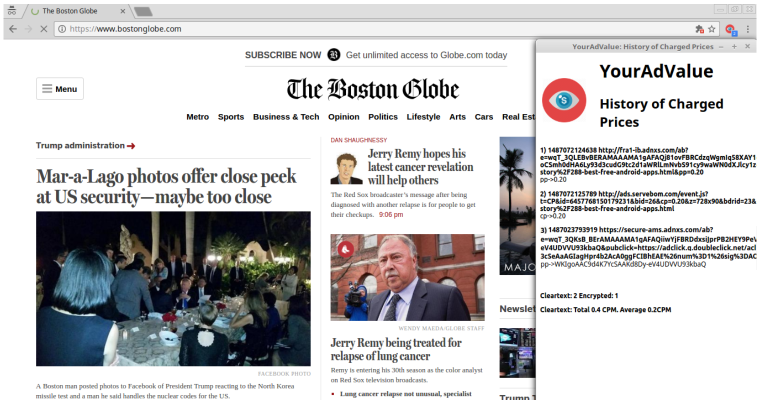
But how do all these trackers monetize from these data? The answer is in the advertising ecosystem, where advertisers are purchasing audiences to deliver their ad-impressions. Therefore, there are studies focusing solely on privacy preservation in the advertising ecosystem. For example, Privad [32] is designed to conceal user activities from an ad-network, by interposing an anonymizing proxy between the browser and the ad-network, allowing a trusted client software to select relevant ads locally. Unfortunately, it requires broad adoption of high-performance anonymizing proxies.

Alternatively, Adnostic [78] is an architecture for interest-targeted advertising without tracking. Like Privad, Adnostic uses client-based functionality to perform ad selection, but eliminates anonymizing proxies at the cost of less precise ad targeting. In [66], authors propose obfuscation of the user's full identity while browsing the web. This was achieved by introducing Web Identity Translator (WIT) in-between the user's client and the visited websites. Given that advertisers are interested in adjusting their buying strategy at real time, it is unclear if such approaches can be adapted to contemporary technologies such as RTB auctions.

The economics of private data have long been an interesting topic and attracted a considerable body of research either from the user's perspective [2, 11, 70, 74], or the advertiser's perspective [14, 19, 22, 62]. In [2] authors discuss the value of privacy after defining two concepts (i) *Willingness To Pay*: the monetary amount users are willing to pay to protect their privacy, and (ii) *Willingness To Accept*: the compensation that users are willing to accept for their privacy loss. In two user-studies [11, 74] authors measure how much users value their own offline and online personal data, and consequently how much they would sell them to advertisers. In [70], the authors propose "transactional" privacy to allow users to decide what personal information can be released and receive compensation from selling them.

In [62], the authors perform an analysis of cookie matching in association with the RTB advertising. Similar to our approach, they leverage the RTB nURL to observe the charge prices and they conduct a basic study to provide some insights into these prices, by analyzing different user profiles and visiting contexts. Their results confirm that when the users' browsing histories are leaked, the charge prices tend to be increased. Similarly, in [61], the authors propose a transparency enhancing tool showing to the users the RTB charge price every time a RTB auction is performed. Furthermore, they collect profiled and un-profiled data from a browser extension and a crawler respectively, and they compare the RTB prices, the bidding frequency and the inter-relations among ADXs and DSPs. Contrary to our work, both studies use a dataset from (i) a small number of 100 users, (ii) over desktop, (iii) covering only one month, (iv) and based on these data, they estimate the advertising total revenues using only the cleartext prices based on the arbitrary assumption that encrypted and cleartext prices follow the same distributions. Although their results regarding the average prices per ad are comparable to ours (~0.5CPM Vs. ~0.26CPM), they are not equal since their study was conducted on desktop and in 2013, when ad spending in desktop was higher than in mobile [12].

In [22], authors use a dataset of users' HTTP traces and provide rough estimates of the relative value of users by leveraging the suggested bid amounts for the visited websites, based on categories provided by the Google AdWords. FDTV [14] is a plugin to inform users in real-time about the economic value of the personal information associated to their Facebook activity. Although similar to ours, our approach works for all HTTP activity of mobile users.


The Boston Globe
https://www.bostonglobe.com
SUBSCRIBE NOW Get unlimited access to Globe.com today
Menu
The Boston Globe
Metro Sports Business & Tech Opinion Politics Lifestyle Arts Cars Real Est
Trump administration
Mar-a-Lago photos offer close peek at US security—maybe too close
Jerry Remy hopes his latest cancer revelation will help others
Jerry Remy being treated for relapse of lung cancer
YourAdValue: History of Charged Prices
YourAdValue
History of Charged Prices

```
1) 1487072124638 http://fra1-ib.adnxs.com/ab?
e=wqT_3QLEBvBERAMAAAMA1gAFAQj81ovFBRCdzqWgmIq58XAY1
oCSmh0dHA6Ly93d3cudG9tc2d1aWRlLmNvbS91cy9waWN0dXJlcy1z
story%2F288-best-free-android-apps.html&pp=0.20
pp->0.20

2) 1487072125789 http://ads.servebom.com/event.js?
t=CP&id=6457768150179231&bid=26&cp=0.20&z=728x90&bdrid=23&
story%2F288-best-free-android-apps.html
cp->0.20

3) 1487023793919 https://secure-ams.adnxs.com/ab?
e=wqT_3QKsB_BErAMAAAMA1gAFAQiiwYjFBRDdxsiJprPB2HEY9PeV
eV4UDVVU93kbaQ&pubclick=https://adclick.g.doubleclick.net/acl
3cSeAaAGIagHpr4b2AcA0ggFCIBhEAE%26num%3D1%26sig%3DAC
pp->WKIgoAAC9d4K7YcSAAKd8Dy-eV4UDVVU93kbaQ

Cleartext: 2 Encrypted: 1
Cleartext: Total 0.4 CPM. Average 0.2CPM
```

Figure 20: Preliminary implementation of YourAdValue Chrome extension in use.

Furthermore, journalists from Financial Times, created an interactive calculator [19] to explore how valuable specific pieces of user data are for the ad-companies. This calculator is based on the analysis of industry pricing data from a range of sources in the US.

Finally, the rapid growth of RTB auctions has drawn the attention of the research community, which aims to explore the economics of the RTB ad ecosystem. In [83], the authors provide an insight to pricing and an empirical analysis of the technologies involved. They use internal data of an ADX and they study its bidding behaviors and strategies. In [81], the authors propose a winning price predicting mechanism by leveraging machine learning and statistical methods to train a model using the bidding history. Their predicting approach aims to help DSPs fine-tune their bids accordingly. Though such studies help us understand some internal mechanisms of ADXs and DSPs, they are not applicable to our setting as we try to infer the cumulative ad-cost of each user based on user-related features that are measurable from the user's device over time.

8 DISCUSSION & CONCLUSION

Limitations. Our approach, through YourAdValue plugin, monitors the charge prices for each auctioned ad-slot. However, there are several cost models in digital ad-buying. For example, Cost-Per-Impression is where the advertiser pays when an impression is rendered, and Cost-Per-Click is where the advertiser pays only if the impression is rendered and clicked, etc. Given that our study is based on passive measurements, we currently unable to determine the cost model of each auctioned ad-slot. Therefore, we assume all charge prices are under the Cost-Per-Impression model, thus computing the maximum cost advertisers pay for a user.

Computing The financial worth of individuals. Via our methodology, users can estimate, at real time, the cost advertisers pay to reach them. However, this work's important technical contribution, i.e., *how to compute the financial worth of individuals* with a passive measurement method has several applications. Our methodology could provide more transparency on what each type of the users' personal data is worth, and allow users to take advantage of, and (re)negotiate their online value with data hub companies who are interested in investing and innovating in the area of targeted advertising. Also, such companies can use our methodology to assess the costs implied in this area, how to allocated appropriate resources and, even, estimating bidding strategies of competitors. In addition, regulators and policy makers could provide guidelines and laws to users and companies for containing the leakage of users' personal data. Finally, tax auditors could estimate ad-companies' revenues, and detect discrepancies from their tax declarations in an independent and transparent way.

Conclusion. In this study, we aim to enhance transparency in the ad ecosystem, where user's personal data is the most important factor affecting the pricing dynamics. We developed a first of its kind methodology to estimate how much do advertisers pay to reach a user. Our methodology leverages the rapidly growing RTB protocol and the new advertising model of programmatic instantaneous auctions, where the advertisers evaluate the users' collected data at real time and bid for an ad-slot in their display. Our study analyzes the RTB price notifications sent to winning advertising bidders and focuses on the distinction between cleartext and encrypted price notifications and how to estimate the latter. Towards this end, we train a model using as ground truth prices obtained by running our own probing ad-campaigns. We bootstrap and validate our methodology using a year long trace of real user browsing data, as well as two real world ad-campaigns. Finally, we designed YourAdValue: a system to allow users to compute at real time the value advertisers pay to reach them. As future work, we plan to make our prototype (a preliminary version can be seen in Figure 20) available for the community to test and explore its effectiveness with online users.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions and help received during the execution of this project: Prof. Vishal Misra and his team (Columbia Univ.), for a preliminary study as part of an early DTL grant; Prof. Evangelos Markatos (FORTH-ICS) and Dr. Claudio Soriente (Telefonica I+D) for their valuable feedback and comments; Jose Ramon Gomez Utrilla (Telefonica 4th Platform team), for his help with the design and execution of ad-campaigns; Costas Iordanou, for his help with the Google AdWords; Rafa Gross-Brown (DTL) on the design of ads; and Xiaoyuan Yang and Martin Levi Gonzalez (Telefonica Niji team) for providing the mobile data.

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 653449 (project TYPES) and Marie Sklodowska-Curie grant agreement No 690972 (project PROTASIS). The paper reflects only the authors' view and the Agency and the Commission are not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. 2014. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security.*

[2] Alessandro Acquisti, Leslie K John, and George Loewenstein. 2013. What is privacy worth? *The Journal of Legal Studies* (2013).

[3] Paul Barford, Igor Canadi, Darja Krushevskaja, Qiang Ma, and S. Muthukrishnan. 2014. Adscape: Harvesting and Analyzing Online Display Ads. In *Proceedings of the 23rd International Conference on World Wide Web (WWW ’14).* ACM, New York, NY, USA, 597–608. https://doi.org/10.1145/2566486.2567992

[4] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. 2016. Tracing Information Flows Between Ad Exchanges Using Retargeted Ads. In *25th USENIX Security Symposium (USENIX Security 16).* USENIX Association, Austin, TX, 481–496. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/bashir

[5] BDEX - Big Data Exchange. 2015. DMP 2.0 - Introduction of the DXP. http://www.bigdataexchange.com/dmp-2-0-introduction-of-the-dxp/. (2015).

[6] Howard Beales. 2010. The value of behavioral targeting. *Network Advertising Initiative* (2010).

[7] Paul Bernal. 2015. Our web history reveals what we think and do. Shouldnfit that remain private? https://theconversation.com/our-web-history-reveals-what-we-think-and-do-shouldnt-that-remain-private-50289. (2015).

[8] BI Intelligence. 2017. Programmatic advertising is under review. http://www.businessinsider.com/programmatic-advertising-under-review-2017-1. (2017).

[9] BlueKai. 2011. Data Management Platforms Demystified. http://www.bluekai.com/files/DMP_Demystified_Whitepaper_BlueKai.pdf. (2011).

[10] bobzilla, arkasha, uhtu. 2001. WiGLE: Wireless Network Mapping. https://wigle.net/. (2001).

[11] Juan Pablo Carrascal, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira. 2013. Your browsing behavior for a big mac: Economics of personal information online. In *Proceedings of the 22nd international conference on World Wide Web.*

[12] Dave Chaffey. 2016. Mobile Marketing Statistics compilation. http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics/. (2016).

[13] Josh Constine. 2016. Facebook crushes Q2 earnings, hits 1.71B users and record share price. https://techcrunch.com/2016/07/27/facebook-earnings-q2-2016/. (2016).

[14] Angel Cuevas, Ruben Cuevas, Raquel Aparicio, and Jose Gonzalez. 2017. FDVT: Data Valuation Tool for Facebook Users. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI ’17).*

[15] Disconnect. 2011. A faster, safer Internet is one click away. https://disconnect.me/. (2011).

[16] Justin Driskill. 2016. Ad Size Guide. http://theonlineadvertisingguide.com/ad-size-guide/300x250/. (2016).

[17] Peter Eckersley. 2010. How Unique is Your Web Browser?. In *Proceedings of the 10th International Conference on Privacy Enhancing Technologies.*

[18] Hazem Elmeleegy, Yinan Li, Yan Qi, Peter Wilmot, Mingxi Wu, Santanu Kolay, Ali Dasdan, and Songting Chen. 2013. Overview of Turn Data Management Platform for Digital Advertising. *Proc. VLDB Endow.* (2013).

[19] Emily Cadman Emily Steel, Callum Locke and Ben Freese. 2013. How much is your personal data worth? http://www.ft.com/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html. (2013).

[20] Let’s Encrypt. 2017. Percentage of Web Pages Loaded by Firefox Using HTTPS. https://letsencrypt.org/stats/. (2017).

[21] Steven Englehardt and Arvind Narayanan. 2016. Online Tracking: A 1-million-site Measurement and Analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS ’16).* ACM, New York, NY, USA, 1388–1401. https://doi.org/10.1145/2976749.2978313

[22] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. 2013. Follow the Money: Understanding Economics of Online Aggregation and Advertising. In *Proceedings of the 2013 Conference on Internet Measurement Conference.*

[23] Google. 2016. DoubleClick for Publishers ? Small Business. https://www.google.com/doubleclick/publishers/small-business/. (2016).

[24] Google AdSense. 2016. Guide to ad sizes. https://support.google.com/adsense/answer/6002621. (2016).

[25] Google Developers. 2016. Real-Time Bidding Protocol - Basics. https://developers.google.com/ad-exchange/rtb/start. (2016).

[26] Google Developers. 2016. RTB Decrypt Price Confirmations. https://developers.google.com/ad-exchange/rtb/response-guide/decrypt-price. (2016).

[27] Google Inc. 2016. DoubleClick Manager. https://www.doubleclickbygoogle.com/solutions/digital-marketing/bid-manager/. (2016).

[28] Google Inc. 2017. Google AdWords. https://www.google.com/adwords/. (2017).

[29] Annabelle Green. 2016. Customer data collection increased to improve customer experience, research finds. http://business-reporter.co.uk/2016/07/20/customer-data-collection-increased-improve-customer-experience-research-finds/. (2016).

[30] Tren Griffin. 2017. Tren’s Advice for Twitter. https://25iq.com/2017/01/06/trens-advice-for-twitter/. (2017).

[31] Saikat Guha, Bin Cheng, and Paul Francis. 2010. Challenges in Measuring Online Advertising Systems. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC ’10).* ACM, New York, NY, USA, 81–87. https://doi.org/10.1145/1879141.1879152

[32] Saikat Guha, Bin Cheng, and Paul Francis. 2011. Privad: Practical Privacy in Online Advertising. In *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation.*

[33] Beau Hamilton. 2016. Google Has Quietly Dropped Ban On Personally Identifiable Web Tracking. https://tech.slashdot.org/story/16/10/22/008216/google-has-quietly-dropped-ban-on-personally-identifiable-web-tracking. (2016).

[34] William T Harding, Anita J Reed, and Robert L Gray. 2001. Cookies and Web bugs: What they are and how they work together. (2001).

[35] IAB. 2015. OpenRTB API Specification Version 2.4. http://www.iab.com/wp-content/uploads/2016/01/OpenRTB-API-Specification-Version-2-4-DRAFT.pdf. (2015).

[36] IHS Technology. 2015. Paving the way: how online advertising enables the digital economy of the future. https://www.iabeurope.eu/files/9614/4844/3542/IAB_IHS_Euro_Ad_Macro_FINALpdf.pdf. (2015).

[37] Interactive Advertising Bureau. 2015. IAB Tech - Lab Content Taxonomy. https://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag-taxonomy/. (2015).

[38] InvestingAnswers. 2017. Cost Per Thousand (CPM). http://www.investinganswers.com/financial-dictionary/businesses-corporations/cost-thousand-cpm-2917. (2017).

[39] iSixSigma. 2017. How to Determine Sample Size, Determining Sample Size. https://www.isixsigma.com/tools-templates/sampling-data/how-determine-sample-size-determining-sample-size/. (2017).

[40] Surya Mattu Julia Angwin, Terry Parris Jr. 2016. Facebook is quietly buying information from data brokers about its users’ offline lives. http://www.businessinsider.com/facebook-data-brokers-2016-12. (2016).

[41] Kate Kaye. 2009. Nielsen in Pact to Use Offline Data for Online Ad Targeting. https://www.clickz.com/nielsen-in-pact-to-use-offline-data-for-online-ad-targeting/77948/. (2009).

[42] Martin Kihn. 2016. Top 10 Amazing Secrets of DMPs. http://blogs.gartner.com/martin-kihn/top-10-amazing-secrets-of-dmps/. (2016).

[43] Ben Kneen. 2011. SSP to DSP Cookie Syncing Explained. http://www.adopsinsider.com/ad-exchanges/cookie-syncing/. (2011).

[44] Knime. 2015. Seven Techniques for Dimensionality Reduction. https://www.knime.org/blog/seven-techniques-for-data-dimensionality-reduction. (2015).

[45] Know Online Advertising Inc. 2013. Data Management Platform fi?! DMP. http://www.knowonlineadvertising.com/programmatic-buying/data-management-platform-dmp/. (2013).

[46] Know Online Advertising Inc. 2013. Definition of Backfill. http://www.knowonlineadvertising.com/advertisingdictionary/backfill/. (2013).

[47] Georgios Kontaxis, Michalis Polychronakis, Angelos D. Keromytis, and Evangelos P. Markatos. 2012. Privacy-preserving Social Plugins. In *Proceedings of the 21st USENIX Conference on Security Symposium.*

[48] Balachander Krishnamurthy and Craig Wills. 2009. Privacy Diffusion on the Web: A Longitudinal Perspective. In *Proceedings of the 18th International Conference on World Wide Web.*

[49] Steve Kroft. 2014. The Data Brokers: Selling your personal information. http://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/. (2014).

[50] Leading Edge Provider. 2016. Internet Trends, stats & facts in the U.S. and Worldwide 2016. http://www.leadingedgeprovider.com/2016/12/internet-trends-stats-facts-in-the-u-s-and-worldwide-2016/. (2016).

[51] Michael Learmonth. 2009. Online Ad Industry: Advertising Is ’Creepy’. http://adage.com/article/digital/online-ad-industry-advertising-creepy/140840/. (2009).

[52] Christophe Leung, Jingjing Ren, David Choffnes, and Christo Wilson. 2016. Should You Use the App for That?: Comparing the Privacy Implications of App- and Web-based Online Services. In *Proceedings of the 2016 ACM on Internet Measurement Conference.*

[53] Lukasz Olejnik and Claude Castelluccia. 2015. To bid or not to bid? Measuring the value of privacy in RTB. http://lukaszolejnik.com/rtb2.pdf. (2015).

[54] MaxMind Inc. 2017. GeoIP Databases & Services: Industry Leading IP Intelligence. https://www.maxmind.com/en/geoip2-services-and-databases. (2017).

[55] Dan Mitchell. 2007. Online Ads vs. Privacy. http://www.nytimes.com/2007/05/12/technology/12online.html. (2007).

[56] MoPub. 2016. MoPub OpenRTB 2.3 Integration Guide. https://dev.twitter.com/mopub-demand/marketplace-integration/openrtb.

(2016).

[57] MoPub Inc. 2017. Mopub Platform. http://www.mopub.com/platform/. (2017).

[58] Natalie Lynn. 2016. Mobile Web vs. Mobile In-App Advertising: Which Is Best for Your Campaign? https://gimbal.com/mobile-web-vs-mobile-in-app-ads/. (2016).

[59] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. 2015. PriVaricator: Deceiving Fingerprinters with Little White Lies. In *Proceedings of the 24th International Conference on World Wide Web*.

[60] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. 2013. Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting. In *Proceedings of the 2013 IEEE Symposium on Security and Privacy*.

[61] Lukasz Olejnik and Claude Castelluccia. 2015. To bid or not to bid? Measuring the value of privacy in RTB. (2015).

[62] Lukasz Olejnik, Minh-Dung Tran, and Claude Castelluccia. 2014. Selling off User Privacy at Auction. In *21st Annual Network and Distributed System Security Symposium, NDSS, San Diego, California, USA, February 23-26*.

[63] OpenX. 2016. RTB Macros. http://docs.openx.com/Content/demandpartners/rtb_macros.html. (2016).

[64] Elias P. Papadopoulos, Michalis Diamantaris, Panagiotis Papadopoulos, Thanasis Petsas, Sotiris Ioannidis, and Evangelos P. Markatos. 2017. The Long-Standing Privacy Debate: Mobile Websites Vs Mobile Apps. In *Proceedings of the 26th International World Wide Web Conference (WWW'17)*.

[65] Panagiotis Papadopoulos, Antonis Papadogiannakis, Michalis Polychronakis, Apostolis Zarras, Thorsten Holz, and Evangelos P. Markatos. 2013. K-subscription: Privacy-preserving Microblogging Browsing Through Obfuscation. In *Proceedings of the 29th Annual Computer Security Applications Conference*.

[66] Fotios Papaodyssefs, Costas Iordanou, Jeremy Blackburn, Nikolaos Laoutaris, and Konstantina Papagiannaki. 2015. Web Identity Translator: Behavioral Advertising and Identity Privacy with WIT. In *Proceedings of the 14th ACM Workshop on Hot Topics in Networks*.

[67] Patricia Gamer. 2015. Average revenue per user is an important growth driver. http://marketrealist.com/2015/02/average-revenue-per-user-is-an-important-growth-driver/. (2015).

[68] PricewaterhouseCoopers LLP,. 2015. IAB Programmatic Revenue Report 2014 Results. http://www.iab.net/media/file/PwC_IAB_Programmatic_Study.pdf. (2015).

[69] PulsePoint. 2016. RTB Implementation Notes. http://docs.pulsepoint.com/display/BUYER/RTB+Implementation+Notes. (2016).

[70] Christopher Riederer, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, and Pablo Rodriguez. 2011. For Sale : Your Data: By : You. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*.

[71] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and Defending Against Third-party Tracking on the Web. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*.

[72] Sandvine Incorporated. 2016. Sandvine: 70% Of Global Internet Traffic Will Be Encrypted In 2016. https://www.sandvine.com/pr/2016/2/11/sandvine-70-of-global-internet-traffic-will-be-encrypted-in-2016.html. (2016).

[73] Judy Selby. 2016. The Impact of Big Data Decisions on Business Valuations. https://datafloq.com/read/impact-big-data-decisions-business-valuation. (2016).

[74] Jacopo Staiano, Nuria Oliver, Bruno Lepri, Rodrigo de Oliveira, Michele Caraviello, and Nicu Sebe. 2014. Money Walks: A Human-centric Study on the Economics of Personal Mobile Data. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

[75] Statista Inc. 2016. Premium Digital advertising spending worldwide from 2015 to 2020 (in billion U.S. dollars). https://www.statista.com/statistics/237974/online-advertising-spending-worldwide/. (2016).

[76] The Office for Creative Research. 2014. Floodwatch. https://ocr.nyc/user-focused-tools/2014/06/15/floodwatch/. (2014).

[77] The Tor Project, Inc. 2002. Tor Project: Anonymity Online. https://www.torproject.org/. (2002).

[78] Vincent Toubiana, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas. 2010. Adnostic: Privacy preserving targeted advertising. In *Proceedings of the 17th Annual Network and Distributed System Security Symposium (NDSS '10)*.

[79] William Vickrey. 1961. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance* (1961).

[80] Sean Whitbeck. 2015. RTB is Growing Like Mad. Is Your Mobile Marketing Keeping Up? http://liftoff.io/rtb-growing-like-mad-mobile-marketing-keeping/. (2015).

[81] Wush Chi-Hsuan Wu, Mi-Yen Yeh, and Ming-Syan Chen. 2015. Predicting Winning Price in Real Time Bidding with Censored Data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[82] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. 2009. How Much Can Behavioral Targeting Help Online Advertising?. In *Proceedings of the 18th International Conference on World Wide Web*.

[83] Shuai Yuan, Jun Wang, and Xiaoxue Zhao. 2013. Real-time Bidding for Online Advertising: Measurement and Analysis. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*.