Information Market for Web Browsing: Design, Usability and Incremental Adoption

Arash Molavi Kakhki
ThousandEyes
Boston, MA, USA
arash.molavi@gmail.com

Vijay Erramilli
Salesforce
San Francisco, CA, USA
evijay@gmail.com

Phillipa Gill
University of Massachusetts
Amherst, MA, USA
phillipa@cs.umass.edu

Augustin Chaintreau
Columbia University
New York, NY, USA
augustin@cs.columbia.edu

Balachander Krishnamurthy
AT&T Labs - Research
New York, NY, USA
bala@research.att.com

ABSTRACT

Browsing privacy solutions are faced with an uphill battle to deployment. Many operate counter to the economic objectives of popular online services (*e.g.*, by completely blocking ads) and do not provide enough incentive for users who may be subject to performance degradation for deploying them. In this study, we take a step towards realizing a system for online privacy that is mutually beneficial to users and online advertisers: an *information market*. This system not only maintains economic viability for online services, but provides users with financial compensation to encourage them to participate. We prototype and evaluate an information market that provides privacy and revenue to users while preserving and sometimes improving their Web performance. We evaluate feasibility of the market via a one month field study with 63 users and find that users are indeed willing to sell their browsing information. We also use Web traces of millions of users to drive a simulation study to evaluate the system at scale. We find that the system can indeed be profitable to both users and online advertisers.

KEYWORDS

Browsing privacy, Information market

ACM Reference Format:
Arash Molavi Kakhki, Vijay Erramilli, Phillipa Gill, Augustin Chaintreau, and Balachander Krishnamurthy. 2018. Information Market for Web Browsing: Design, Usability and Incremental Adoption. In *Proceedings of (Performance '18).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Most online services currently provide users the same terms: users may use them free of charge but those uses are monetized through data collection used for online advertising. This economic model was long criticized by privacy advocates, due to the growing amount of information collected about each user [29, 34], and the lack of effective control it provides her about how it's used. Indeed, this model recently came under a new stress: Softwares that block ads and/or limit some third party tracking [8, 15, 42] are no more confined to a minority of savvy users: Adoption of those by Internet users is already reported up to 35% in two European countries in 2015, and more importantly, it is quickly growing everywhere, including in the largest advertising markets [11]. Some ad-blocking even comes shipped for default browser on mobile platforms such as Apple iOS 9 [1]. Those users who *in effect* opt-out of tracking or advertising (at least partially) are estimated to negatively affect ad revenues in the tens of billions [11]. Those revenue losses motivated multiple research prototypes and development for alternative ways to collect, manage and exploit personal data: personal data store, lockers, intentcasting and privacy preserving ad personalization, see [9, 21, 41] and more than 50 related development efforts mentioned at cyber.law.harvard.edu/projectvrm/.

Online privacy solution, like those mentioned above, typically struggle with a restricted deployment due to the incentives (or lack thereof) that they produce. On the one hand a unilateral opt-out solution - deployed by a user herself, such as blocking [15, 42] - may be difficult to configure as it often impacts service quality [25]. There is even an incentive for ad-networks, aggregators, and publishers to cause disruption in service for such opt-out users, as its adoption by a user reduces the value she generates significantly [17, 28]. In contrast, a cooperative solution - deployed jointly by multiple parties (*e.g.,* aggregators, publishers, users) [30, 49] - may in principle offer incentives to all. However, most of the systems aforementioned offer users enhanced privacy but no more, and little is formally known about the incentives offered to other parties. Many evidence suggests that enhanced privacy alone can be difficult for users to perceive and treat as a rational choice, even among users self-reporting a concern for tracking [14] (a trend we confirm in our experiment).

Here we evaluate through experiments and data-driven analysis the promise of a different online privacy solution: *an information market*. In contrast to all systems above, an information market implements an incentive for all parties to participate. Users, for instance, can enjoy enhanced privacy by selecting which of their data can be used and be compensated for those. Depending on how the information market is designed, other parties may find it profitable to use it for their interaction with online users. The idea of information markets is not new [20, 27, 36, 45], even offered by some

Performance '18, December 5-7, 2018, Toulouse, France
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

products today [2–4][1] but very little is known of their economic viability. Even less is known about how they should be designed to engage users, let alone how they could gradually encourage an incremental adoption. This paper provides a much needed evaluation of each of those aspects, concluding that information market show some promise. While information markets could be applied to multiple forms of personal data, in this work we focus on their application to third party tracking during web browsing, for two reasons: First, third party tracking is a critical component of today's online advertising; second the sharing of browsing data with third parties is often cited first among causes of concern by online users [12].

This paper presents the following contributions

- We design and build a simplified information market for third party web tracking. Through a simple architecture, this system enables selective privacy protection and economic transactions over data. Our design seamlessly integrates with today's web tracking and ad-network functions to be backward compatible. In fact, it even allows users to benefit from a performance boost as we show it can easily be combined with web acceleration features. (§3)
- We conduct a 30-day experiment with 63 participants in two metropolitan areas to observe the effect of economic incentive on online users' behaviors. We confirm that self-reported privacy attitudes may not always be aligned with actual data disclosures by users. More importantly, we validate that the two features of an information market, *i.e.,* privacy when you want it, money when you do not, are effectively used and managed by users. Every user has some data they choose to protect, although disclosing it would increase their earning up to 52%. As their browsing expands and new data gets created, users also often choose economic returns over data protection. In fact, we observe a growing engagement during our user study, and an overwhelming majority claims to be likely to use such a system if it was deployed. (§4)
- We study the potential for an incremental deployment of an information market to a large scale. To do so, we extend a model predicting today's online advertising revenue per user. Through cooperative game theory we analyze how market forces may affect revenue redistribution to different parties should information market be offered as an option. We then assume that a party adopts an information market only when it gains positive benefits from it, which may in turn affect others' decision to adopt. We use this model to analyze spread of adoption in several HTTP traces containing up to 3 million users, along with the publishers and third parties they interact with. This model predicts significant revenue growth (up to 9-12%) as information market expands the availability of high quality data about users. A significant fraction of the users (from 35% to 92% depending on the traces) adopt information market and receive monetary gain. (§5)

[1]For simplicity we focus on financial incentives here, but compensation can be in the forms of upgrade or discounts.

These results present to our knowledge the first data driven evaluation of information markets and their potential to scale to the web. We recognize that deployment of an information market remains a hard problem: Online user privacy is constantly getting more complicated, as new forms of tracking or re-identification [13] come into play. The complexity of the online advertising ecosystem creates multiple frictions to deployment[50]. Moreover, there is economic incentives for users to game the system (we monitored our user study closely and confirm we did not observer any instances of gaming to pollute our data and findings, and discuss in §4.3 how gaming can be detected). Given those limitations, it is important to interpret our results in their context: First, they assume for the information market to operate such that users can technically opt-out. Also our architecture allows the privacy preserving module to evolve to handle new forms of tracking and re-identification. Second, our analysis of online revenue and its redistribution necessarily makes some assumptions (following bargaining resolution as classically drawn by Nash and later Shapley). Our results prove at least that a few conventional wisdoms about online privacy may not always hold. Solutions offering privacy choices and compensating users for data may not necessarily reduce overall advertising revenue; they may even benefit publishers. This contributes to the ongoing debate over who benefits from targeted advertising [39]. Finally, in terms of design, our results concur to prove the promise of enabling selective privacy: our experiments show that users are able to determine the data they want to protect, and those they are eager to sell. For the latter, our trace analysis revealed for the first time that opportunities abound for an information market to make data more widely available, fueling more revenue. These trends may be reproduced under different conditions and assumptions than ours. Our findings contribute to argue that the arm race we experience between blocking and more invasive tracking techniques may not necessarily in the long term serve the interests of online publishers and the advertising industry overall.

2 BACKGROUND AND REQUIREMENTS

Targeted advertising has increased in usage over the last few years and generally comes in multiple varieties: contextual, retargeting and behavioral [38]. The latter is offered by Google since 2009 [5]. While contextual advertising serve ads based on the content of the page embedding the ads, both retargeting and behavioral utilize browsing history of a user to place relevant ads.

Past browsing behavior is obtained by aggregators via embedding themselves on Web pages as *3rd-parties* in combination with setting cookies in the browser (NB: Other techniques exist such as fingerprinting we discuss them more in Sec. 3.1). Consider the following example: (i) user Alice visits publisher pubA.com which contains references to a 3rd-party aggregator: agg.com. (ii) If Alice is visiting pubA.com for the first time, the HTTP response is the content of the page being requested along with a Set-Cookie HTTP header, with cookies pertaining to pubA.com. Likewise, if Alice has never visited agg.com (or any page with them as a 3rd-party), the responses from this domain will also include Set-Cookie header. (iii) Alice next visits publisher pubB.com that also contains references to aggregator agg.com. As a cookie for agg.com

was set when she visited pubA.com, Alice's browser sends this cookie to agg.com along with its request. Thus, agg.com now knows that Alice visited both pubA.com, and pubB.com and can use this to customize which ads to show.

This type of 3rd-party tracking is considered objectionable by many, because it results in Alice's browsing history being revealed to 3rd-parties that she is unaware of. While publishers themselves may also track Alice over time – 1st-party tracking – Alice is generally aware of 1st-parties that she is dealing with. In contrast, a publisher may embed any number of 3rd-parties without notifying the user. *The first goal of our system is to give transparency and control to Alice in the context of 3rd-party tracking.* A key difference with previous solutions to that problem is that Alice - aware that part of her browsing history may boost the revenue of the ads shown on the websites she visits - may sometimes be eager to disclose some of that information *selectively* for an appropriate reward. *Permitting such data transaction is the second goal of our system.*

Overall, we hope to design an information system that can be built and used experimentally. It should hence satisfy the following requirements:

(1) **Selective Privacy protection.** The system should protect users' browsing history from being revealed to 3rd-party aggregators. Note this protection is critical even for the data the user intend to sell. Without privacy protection there is no incentive for aggregators to enter the market as they can obtain users' data via conventional means. The system should generally enable users to disclose only a fine-grained subset of their browsing.
(2) **Backwards and incentive compatibility.** First, this means that the system should work with today's tracking and online advertising systems, with minimal modifications required for the delivery of ads. This helps in reducing friction in adoption for data aggregators and ad-networks. Second, it should encourage adoption by providing different parties with the right economic incentives. Note that this last requirement is more complex as decisions of several parties together interact to change revenue. This is why we will carefully study how different designs affect online revenue sharing.
(3) **Access to data.** Unlike existing proposals [19, 30, 49] an information market should not require aggregators to communicate targeting algorithms–which may be considered trade secrets–to the system. Instead, aggregators should be able to purchase raw data about users' browsing habits (*i.e.,* when and what sites the user has visited).
(4) **Avoid having users price data.** Since aggregators have knowledge of the relative value of user data (*e.g.,* somebody who visits Rolex.com may have higher value) they are in the best position to determine prices in the market. Further, having users assign value to their own data [18] or calculate the loss of utility with their information release [26] is non-trivial. We design information market where price is set by the demand for data from aggregators, we believe those are in a better position to price the data appropriately.

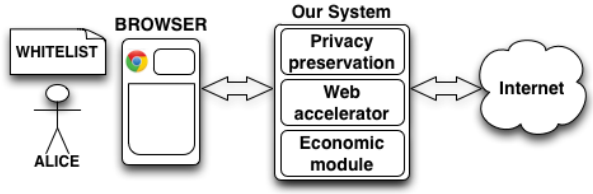

Figure 1: Overview of our system. Alice opts in and creates a whitelist, comprising of sites she's willing to disclose her visit to. Alice's Web requests are routed thru our system. See Sections 3.1-3.3 for detailed description for each module shown.

Figure 2: Sample of whitelists along with visit frequencies shown to aggregators to valuate users and bid on them. Note that aggregators will not know who users 1 and 2 are until they bid on them and win the auction on these users. They are then granted *access* to these users and can track them on their whitelisted sites.

3 SYSTEM DESCRIPTION

Our market solution, designed for the aforementioned requirements, is deployed in the network by a trusted third party *e.g.,* a government body or an ISP, as illustrated in Fig. 1. First, we implement a privacy preservation module to keep user's browsing history private from 3rd-party aggregators. Users who opt-in to the market have all their Web requests routed thru our system, which provides privacy protection by blocking known forms of Web tracking (Sec. 3.1). Note that we do not block ads (or any other requests) and allow them to be shown to the user. The system only prevents leakage of personally identifiable information (PII) to advertisers.

Each user then creates a *whitelist*, which is a list of Web sites, if any, they are willing to be tracked on by 3rd-party aggregators in return for monetary compensation. Users' whitelists along with their frequency of visits to those sites are then anonymized and presented to participating aggregators. Fig. 2 shows a sample of what aggregators see. Aggregators can then valuate each anonymized user based on their whiteliste and frequency of visits to these sites, and bid on them if interested (Apx. B). The system then runs an auction (Sec. 3.3) for each user to determine the winning aggregators, if any, and the winning bid. Each winning aggregator then pays that user the winning amount and in return gets *access* to the user for the duration of the next auction period. We define access as the ability to track the user across their whitelisted sites and use this information as input to any proprietary targeting algorithms to serve targeted ads to the user on those sites.. Note that adding

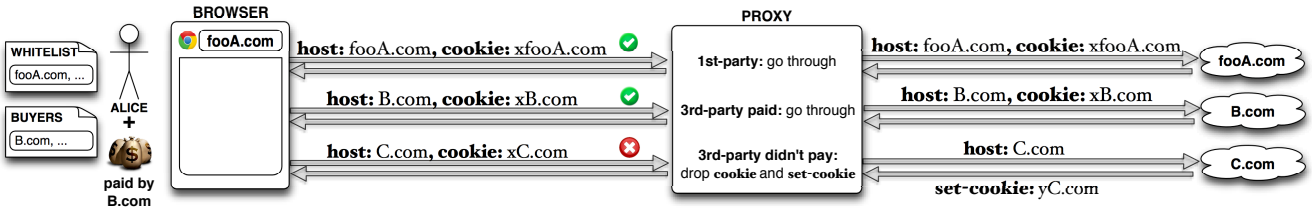

Figure 3: Alice visits a whitelisted site fooA.com **with two trackers** B.com **and** C.com**. Since** B.com **has paid to access Alice, it receives cookies.** C.com**'s cookies are dropped in both directions. Note** B.com**'s cookies would be dropped if** fooA.com **in not whitelisted.**

a Web site to the whitelist by itself does not allow aggregators to track the user on that site. Aggregators still have to bid, win, and compensate the user before they can track them.

Next, we discuss each module in more details.

(1) *Privacy preserving module:* handles all web requests of all users and provides privacy.
(2) *Web acceleration module:* compensates for any performance loss due to the added privacy protection.
(3) *Economic module* handles auctions and economic transactions between the users and aggregators.

3.1 Privacy preserving module

The privacy preserving module handles all HTTP requests and responses for users with the objective to provide privacy protection against 3rd-party tracking, *without* blocking any requests, including advertisements. Below we discuss different tracking methods and how the privacy preserving module deals with each of them:

Cookies: HTTP cookies are the most prevalent method for online tracking. Using cookies, a tracker can assign a unique identifier to each user, which is consistent across different websites, and gives them the power to track users browsing behavior. Today's browsers ship with options to block 3rd-party cookies. However, blocking 3rd-party cookies will not necessarily prevent 3rd-party tracking since most browsers only block Set-Cookie in HTTP responses, meaning if a 3rd-party already has its respective cookies set, *e.g.,* via popups as a 1st-party, then these previously set cookies will indeed be sent out by the browser to the 3rd-party tracker, defeating the purpose [46]. To prevent unwanted 3rd-party tracking via cookies, we first classify each HTTP request as either *1st-party-request* or *3rd-party-request.* A HTTP request is a 1st-party-request if the Host and Referer headers belong to the same root domain, and a 3rd-party-request otherwise [2]. Next, if a request is marked as a 3rd-party-request, we will only let its cookies go through along with the request if both of the following conditions are satisfied:

(a) the user making this request has whitelisted the root domain of the Referer,*i.e.,* the top domain the user is visiting in their browser,
(b) the root domain of the Host, *i.e.,* the 3rd-party destination of the request, has paid to access this user (Sec. 3.3).

Cookies are stripped off from the 3rd-party-request and its response otherwise (Fig. 3). Set-Cookie per-se does not leak any PII, but it is dropped to preserve user's cookie jar consistency. Note since we only strip information off of requests and don't block them, ads would still be rendered properly in the browser.

Referers: Referer headers may contain unnecessary PII [35], and we treat them the same way as cookies, *i.e.,* drop if 3rd-party-request and one or both of the conditions mentioned above are not satisfied.

Web bugs: 1x1 pixel bugs on a page are invisible to the user and serve no content. Their sole purpose is user tracking. The prior steps render these useless.

Etag **and** If-None-Match**:** These optional headers are generally used for web cache validation but can also be used for tracking [35]. When a web server is responding to a request, it can send an Etag header that is an identifier for the version of the resource being requested. When the user makes the request to the same resource in the future, she can send a conditional request using the If-None-Match header. Instead of resource identifier, Etag can contain a unique user identifier for all new users, *i.e.,* users with no If-None-Match header in their request. When the same user returns to the web server, they identify themselves by sending the unique identifier as If-None-Match header, hence enabling tracking. We drop Etag and If-None-Match headers[3].

Flash cookies and LocalStorage: Flash cookies and LocalStorage can be used to respawn deleted cookies [16]. However the way we deal with 3rd-party cookies, renders respawning useless.

JavaScript: JS can be used instead of Set-Cookie header. We already deal with this as we deal with cookies. JS can be used by 3rd-parties to collect information about users, using techniques such as fingerprinting [10]. Such code can be fingerprinted and blocked. This may be included in future versions of the system, but is beyond the scope of this paper.

3.2 Web acceleration module

To mitigate overheads of the privacy protection, we include a Web acceleration module. To boost performance we use standard acceleration methods such as performing on-the-fly prefetching of

[2]With some exception cases, such as CDNs, which are treated as 1st-party.

[3]The web acceleration module (Sec 3.2) mitigates any possible bandwidth overhead to the web server.

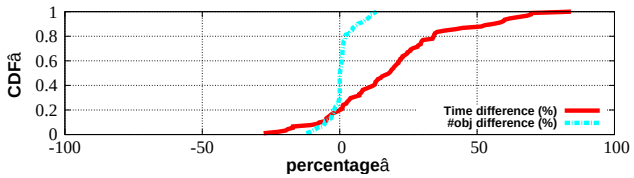

Figure 4: Performance and functionality of web acceleration and privacy preservation module. Positive means we're doing better.

static objects (images, JS, css), prioritizing traffic, image compression, and maintaining persistent TCP connections with both the user and the origin server and using the same connection for multiple requests.

Benchmarks: In order to test our web acceleration and privacy preserving modules for performance and functionality, we wrote a script using PhantomJS [6] and ySlow [7] to load each Alexa top 100 sites once via our system and once directly. We repeat this test 10 times and look at the average difference in load time and number of objects as a measure of functionality[4]. Fig. 4 shows that roughly 80% of sites load faster with our accelerator. With regards to the number of objects, while most sites have exact same number, some have less or more objects when loaded via our system. This is an expected result since most pages are dynamic and the content does not necessarily remain the same over the course of the experiment. We manually inspected pages with different number of objects to confirm that the difference is mostly due to content change.

3.3 Economic module

At the beginning of each auction period, the economic module does the following:

(1) Presents aggregators with *anonymized* whitelists and frequency of visits.
(2) Collects aggregators bids on users. Aggregators bid independent from each other.
(3) Runs an auction per user and determines the winning value and winning aggregators (if any) for each of the users. Note that a user's browsing behavior can be sold to multiple aggregators.
(4) Handles transactions between users and aggregators and makes sure every aggregator pays all the users they won access to. It then notifies the privacy preserving module to grant appropriate access for the duration of the next auction period.

We let aggregators compute the value of access to a user for the following reasons. First, aggregators have experience deriving value from PII. Second, they are able to assess revenues on a short-term basis through the sale of goods or ad-space, compared to the long-term risk a user must calculate in dealing with privacy. Finally, aggregators deal with many customers, and can take more risk in over- or underestimating the value of access, as opposed to users who are more risk averse. However, there can be strong incentives for aggregators to lie about their valuation by under bidding, which results in lowering user's revenue. To prevent such behavior, the auction should have a truth telling mechanism where aggregators gain no additional benefit by under bidding. For this purpose, we rely on an auction mechanism called the exponential mechanism [40] that has truth telling properties, and has been shown to be close to optimal in terms of revenue for the seller (user in our case)[5]. Apx. A explains the auction mechanism in detail.

[4]The expectation is that privacy protection mechanisms do not block any legitimate content that can lead to lower quality of experience for the end-users.

4 USER STUDY

We implemented our system and performed a field study to understand user engagement, usability, and how users perceive economic rewards for their data.

4.1 Implementation

The privacy preserving and web acceleration modules are managed by a proxy, hosted on Amazon EC2, that handles all requests and treats 3rd-party-requests according to the mechanisms described in 3.1. For each user, the proxy maintains a whitelist, as well as a *buyers list* that includes aggregators who have paid to access that user.

We developed a browser extension to serve as an interface for users to interact with the market. Figure 5 shows a screenshot of the extension. It places a color-coded icon in the address bar; green for whitelisted sites, red otherwise. Users can click to add/remove sites to/from the whitelist, or check their current whitelist, buyers list, and past earning.

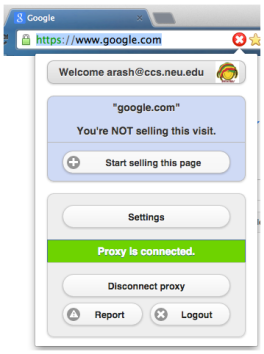

Figure 5: Screenshot of our browser extension. User is visiting google.com **which is not in their whitelist.**

4.2 Experimental setting

User study description A total of 63 participants located in New York City and Boston used our system for 30 days. Our population overrepresents male (75%), and young (85% under 25) users, and has wide variations in income (median around $35k), technical proficiency and Internet use.

Every user was initially assigned a default whitelist containing 108 websites chosen from Alexa top 125 sites excluding social network and adult sites. Users could modify their whitelist at any time. Auctions, using the mechanism shown in Sec. 3.3, were conducted every 3 days, resulting in 10 data points per user. At each auction,

[5]We choose this latter objective, while noting that other objective functions (*e.g.*, maximizing revenue for all players in the value chain) can be chosen.

	Question	Answers				
		5	4	3	2	1
Pre	**1-Are you concerned about protection of your private data online?***	21%	19%	47%	11%	2%
	2-Do you read the privacy policies of the websites you visit?*	0%	3%	27%	37%	33%
	3-Do you use private browsing/Incognito mode while browsing?*	0%	14%	36%	40%	10%
	4-Do you let mobile applications access your location on smartphone?*	5%	25%	36%	27%	7%
	5-Do you maintain several passwords for multiple websites?*	13%	41%	38%	8%	0%
	6-Do you think ads you see online are becoming more relevant to you?*	14%	27%	35%	21%	3%
	7-How much are you willing pay for privacy protection?**	2%	3%	14%	49%	32%
	8-How much do you value today's free online services?**	16%	14%	29%	35%	6%
Post	**1-How comfortable were you to receive micro-payments?***	39%	34%	17%	8%	2%
	2-How satisfied are you with the amount you got for you information?*	17%	50%	18%	13%	2%
	3-How satisfied are you with the performance of the system?*	20%	45%	27%	5%	3%
	4-Did system give you better transparency on which data about you is used?***	57%	32%	7%	3%	1%
	5-Are you likely to use such a system if it was offered large scale?*	56%	20%	15%	7%	2%
	6-Was the system fair in recognizing the value of your information?***	38%	42%	12%	8%	0%
	7-Did such a system increase your concern about your personal data?*	13%	20%	27%	23%	17%
	9-Did system increase awareness on how online services monetize your data?*	28%	38%	17%	12%	5%
	10-Are you likely to pay for privacy now that study has ended?**	0%	0%	28%	33%	39%
	11-Did using this system make you more likely to choose a paid service that comes with privacy guarantee over a free service?*	3%	7%	12%	35%	43%
	12-How likely are you to pay for the online services that you use for free today after this study?*	3%	7%	15%	35%	40%

Table 1: Questions asked in our pre and post study questionnaires. Users answer on a 5-point likert scale: *(5: A lot, 1: Not at all), **(5: >$50, 4: $20-$50, 3: $5-$20, 2: $0-$5, 1: $0), *(5: completely agree, 1: completely disagree)**

the goods for sale are the ability to track users for the upcoming period. Each winning aggregator pays the winning bid value to the user and gains access to her browsing activity *on her whitelisted sites only* during the subsequent auction period to deliver targeted ads. As an example, aggregator agg.com wins access to user u with whitelist WL_u. This implies that, until the next auction, the privacy preserving module will NOT modify 3rd-party-requests to agg.com caused by u while visiting sites in WL_u.

We did not have access to real aggregators for our study, hence we *emulated* 128 aggregators, including big players such as doubleclick and Facebook. The challenge when emulating aggregators is computing their valuation for users' data. For this we use a simple model based on keywords and their estimated cost-per-click commonly made available by ad-networks (details in Apx. B). Note that, based on the emulated valuation of their data and the auction's results, users were paid using *real* money.

Each user was also asked to answer a questionnaire before and after the study, The pre-study questionnaire included demographical questions as well as self reported assessment of user's view on online privacy. The post-study questionnaire contained questions about user's experience with the system, and assessed if using our system changed her view on online privacy. Table 1 shows the questions and answers we received.

4.3 Users and information markets

We made the following observations during the user study:

i) Users actively engage with the system and edit their whitelists. We first studied users' engagement with the tool and how they reacted to different facets of the system. We logged the number of HTTP requests coming from each user, which we refer to as *activity*. The average activity stays in the [10k,14k] range per user during all 10 auction periods, which indicates users were using the system on a regular basis. Users actively added/removed sites to/from their whitelists, showing their engagement with the system, and we observed a sustained growth of the whitelists used, with an average user adding overall 50 new sites (a net 45% increase from default whitelist) during the experiment.

Bucket	earning (USD)	size of WL*	#WL visits	WL/ALL activity
HD	2.76	230.6	151.6	0.71
MD	2.12	128	128.1	0.52
LD	1.55	115.1	67.4	0.48

Table 2: averages of earning per auction, size of whitelist, number of whitelisted visits during auction period, and the fraction of activity which is whitelisted for different buckets. *WL = Whitelist

As the post-study responses suggest Table 1, the majority of users were happy with system's performance, receiving micro-payments, and the amount they were making for the data they released.

ii) Users' disclosure does not match up with their stated privacy attitudes. We looked into users' behavior with regards to data disclosure and how it correlates with their *self-reported* privacy concerns . We took the sum of the 5-point Likert scale answers that users' gave to privacy related questions in the pre-study survey. This gives a *privacy score* for each user, ranging from 8 to 20, where higher scores indicates higher privacy concerns. Surprisingly, we found *no* sign of negative correlation between users' privacy score and size of their whitelist; the associated correlation coefficient is low: -0.1. This, along with previous results [14], suggests that users stated privacy concerns are at odds with their behavior. One possible explanation for this is that users are swayed by the economic returns.

iii) Users trade off their revenues to preserve privacy of their browsing behavior. To understand earning potentials of different users, we sort users based on their whitelists' size and split them into three equal size buckets: high, moderate, and low volume disclosers (resp. HD, MD and LD). Table 2 shows earning, whitelist size, number of whitelisted visits, and fraction of whitelisted activity for each buckets, averaged across users and auction periods. HD users earn 31%/79% more than MD/LD users. A primary factor is that they have larger whitelists (80%/100% more than MD/LD) that results in 71% of their activities coming from whitelisted sites, in contrast with 52%/48% for MD/HD users. If users were to disclose everything, *i.e.,* add *all* the sites they browse to their whitelist, HD, MD, and LD users could potentially increase their earnings by 29%, 48%, and 52% respectively. In other words, while users' behavior may not exactly follow their initial self-reported concerns, they are willing to trade off their revenues to preserve privacy of their browsing behavior.

iv) Users pay attention to their earning and have incentive to game the system. As there are economic incentives in play, it is only natural to expect gaming by users – users may decide to expand their whitelists and browse more in order to increase their economic returns. A key challenge is to differentiate between users who are abusing the system and those who are genuinely releasing more data, by choosing to be less private and/or are more active than average users.

To study this aspect, we sorted users based on their total earning in auctions 1 to 5, and divided them into three groups: top-earners (8 users), low-earners (8 users), and avg-earners (rest). On average top-earners made 6X more money than low-earners and we observed, as expected, top-earners have indeed bigger whitelists (36% bigger than low-earners on average) and browse more (4.5X more than low-earners on average). To evaluate whether users would adjust their behavior and 'game' the system to increase their earnings, for the rest of the auctions, *i.e.,* auctions 6 -10, we gave top-earners a third of their earning and low-earners 3 times their earning. While this change did not have any significant effect on low-earners' behavior – pointing to normal behavior – we did make a few interesting observations about top-earners: 1) we received an email from one of the top earners complaining that the earning suddenly dropped even though their browsing habits had not changed 2) two top earners completely abandoned the system after the 7th auction, 3) three top-earners roughly doubled their whitelist trying to lift up their earning back to previous range, 4) one top earner doubled his/her browsing activity. This suggests users pay attention to their earning, especially once it is significant, and that there is room for gaming the system. Our experiment did not prevent gaming, it is however feasible in the future to include anomaly detection methods to detect abnormal activities.

v) Information market impacts privacy attitudes. Most users agree after the experiment that the system gives them better transparency (96%), increased their awareness of data monetization (83%), and they claim to be likely to use it if it is deployed (91%); see Table 1. The experiment did not noticeably increase their privacy concerns, nor did it affect substantially the price they would pay for privacy. In other words, while users may still not pay for privacy, information markets did help them form a more consistent opinion about their online privacy and value of their data.

5 INCENTIVES FOR ADOPTION

An information market affects the revenue produced by advertisement overall and also how it is shared. We capture this effect first in a model focusing on the transaction associated with a single user. We then describe data used to feed this model for millions of users, and a simple adoption dynamics based on immediate financial incentive. Putting those together, we can then study the spread of adoption and the effect of information market at web scale.

5.1 Incentive for a single user

Our model leverages recent modeling of online advertising [29] to understand how advertising revenue changes with different availability of information. Understanding how revenue is to be distributed among all parties is more difficult to predict. But one way is to apply the theory of cooperative games [48] which offers a principled approach to compute the expected outcome of a negotation in the presence of alternative offers.

5.1.1 Advertising revenue model and assumptions. Our model is inspired from prior work [29] breaking down the value of ad-impression, or cost-per-mille (CPM), into a function of three factors:

$$\mathrm{CPM}(u,p,a) = \mathrm{RON}_a \times \mathrm{TQM}_p \times \mathrm{Int}_a(u)\,, \tag{1}$$

where RON_a (run-on-network) is an ad's nominal cost in ad network a, TQM_p (traffic quality multiplier) captures the quality of the publisher p (*e.g.,* reputable publishers vs. sites distributing copyright infringing content), and $\mathrm{Int}_a(u)$ (intent) depends on the value of information gleaned about the user u by the ad network. Intent captures the fact that an impression can be sold for a larger price if the ad network knows that the user has performed some previous actions (*e.g.,* frequent or recent visits to a product related webpage).

We include three cases for intent. 1) If the user chooses not to reveal any information, this coefficient is by convention equal to 1. 2) When tracking is not blocked, it takes on the value which we refer to as *implicit* intent, $\mathrm{Impl}_a(u) \geq 1$, which depends on how much information the ad-network can collect about the user's browsing. 3) If the user decides to release *all* legitimate sites available to a, the impression can be sold with a higher *explicit* intent $\mathrm{Expl}(u) \geq \mathrm{Impl}_a(u)$ independent of a.

Revenue sharing. We treat the advertising transaction as a cooperative game capturing a simple dynamic: the more user information is known by the ad-network, if they cooperate via an information market, the larger the revenue. We assume that each player is incentivized by receiving a share of the revenue computed using the *Shapley value* [48]. This mechanism has two advantages: it ensures some form of fairness, and maximizes the likeness of cooperation, *i.e.,* users and aggregators participating in the market (see Apx. C an overview of the Shapley value).

In practice, the aggregator a typically collects payments from advertisers, gives a constant fraction $(1-\alpha)$ to the publisher, and then pays for data on the market at price set by this user's Shapley value. The market provider also receives its Shapley value from a.

Mediated marketplace

$\mathcal{S}$	$\text{Int}_a(u)$
$\emptyset$, $\{u\}$, $\{a\}$, $\{m\}$,	$\text{Impl}_a(u)$
$\{u, a\}$, $\{a, m\}$	$\text{Impl}_a(u)$
$\{u, m\}$	1
$\{u, a, m\}$	$\text{Expl}(u)$

	Shapley value
u, m	$\frac{\text{Expl}(u)-1-\frac{3}{2}(\text{Impl}_a(u)-1)}{3}$
a	$\text{Impl}_a(u) + \frac{\text{Expl}(u)-1}{3}$

Direct marketplace

$\mathcal{S}$	$\text{Int}_a(u)$
$\emptyset$	$\text{Impl}_a(u)$
$\{u\}$	1
$\{a\}$	$\text{Impl}_a(u)$
$\{u, a\}$	$\text{Expl}(u)$

	Shapley value
u	$\frac{\text{Expl}(u)-1-2(\text{Impl}_a(u)-1)}{2}$
a	$\text{Impl}_a(u) + \frac{\text{Expl}(u)-1}{2}$

Publ. DNT mediated

$\mathcal{S}$	$\text{Int}_a(u)$
$\emptyset$, $\{u\}$, $\{a\}$, $\{m\}$,	1
$\{u, a\}$, $\{a, m\}$	1
$\{u, m\}$	1
$\{u, a, m\}$	$\text{Expl}(u)$

	Shapley value
u, m	$\frac{\text{Expl}(u)-1}{3}$
a	$\frac{\text{Expl}(u)-1}{3}$

Publ. DNT direct

$\mathcal{S}$	$\text{Int}_a(u)$
$\emptyset$	1
$\{u\}$	1
$\{a\}$	1
$\{u, a\}$	$\text{Expl}(u)$

	Shapley value
u	$\frac{\text{Expl}(u)-1}{2}$
a	$\frac{\text{Expl}(u)-1}{2}$

Table 3: Various market types and coalition scenarios: coefficient multiplying advertising revenue for each coalition scenario (top), Shapley value for each party (bottom).

5.1.2 Allocation of revenue to each player. In this section we investigate the impacts on the Shapley value of different cooperative scenarios between players, *i.e.,* users, aggregators, and the market, under various market implementations. Table 3 shows the intent, $\text{Int}_a(u)$, which impacts the revenue that different cooperating subsets of players, S, can produce.

We start with the situation in which a **mediated market** m centralizes data collection and/or analytics as in [31, 44] (leftmost subtable in Table 3). Without any cooperation, as in today's status quo, a user is tracked implicity, $\text{Int}_a(u) = \text{Impl}_a(u)$. However, if user and aggregator join the market, a higher revenue can be obtained corresponding to $\text{Int}_a(u) = \text{Expl}_a(u)$. In one case, assuming m and u collaborate (*i.e.,* user sells her data) but a does not (*i.e., a* decides not to buy it), revenue effectively decreases as implicit tracking is now blocked, $\text{Int}_a(u) = 1$.

The analysis of this game yields Shapley values for each player (shown in Table 3 bottom). These values specify how much each player receives from the *surplus*, *i.e.,* the potential revenue that is not produced today due to lack of cooperation, which is proportional to the added value of $\text{Expl}(u)$ over $\text{Impl}_a(u)$. Based on Table 3 we make the following observations:

i) Ad-networks are always better off buying the users' data. Buying data makes them join the coalition, always leading to a revenue increase. Hence they always receive a positive Shapley value (Table 3).

ii) Users are not always better off selling their data. Precisely, a user is able to claim a positive share of the surplus only if this surplus is large enough to compensate for the blocking of implicit tracking. In a mediated marker, it occurs as $(\text{Expl}(u) - 1) - \frac{3}{2}(\text{Impl}_a(u) - 1) > 0$ or $r_a(u) > \frac{3}{2}$, where $r_a(u) = \frac{\text{Expl}(u)-1}{\text{Impl}_a(u)-1}$ is the *consent tracking lift*.

iii) Mediated markets lower the bar for users to gain revenue from the market compared with direct markets. Our analysis extends to different market implementations. While mediated markets simplify deployment, and offers a single point of sale to aggregators, other architectures run data collection and/or analytics on users' end [23, 49]. This creates a **direct market** with aggregators entering transactions with users directly (Table 3 second from the left). All previous observations generalize except that users are better off only if $r_a(u) > 2$.

iv) Publishers enforcing Do-Not-Track create ideal conditions to deploy an information market . We also analyze information markets deployed over sites that enforce Do Not Track (Table 3, third and fourth from left). In these scenarios publishers comply with regulations or aim to improve their image vis-a-vis privacy watchdogs and aggregators *and* users are *always* better off buying/selling data via the market.

v) As soon as the possibility of users adopting an information market is credible, a publisher always benefit in the long term from enforcing DNT . Note that publishers are driven by revenue, and since they receive a fixed share, they have an incentive to keep intent (and thus revenue) as high as possible. Enforcing DNT might initially undermine publishers' profit, however it does incentivize users to sell their data, making publishers' revenue higher in the long term.

Information markets increase ad-revenue and redistribute part of the surplus to all participating entities. The price of data set by the market always ensures that aggregators make a profit. However, it does not necessarily mean that a user is always better off selling their data. In addition, a user's decision to join a market is done only once, affecting all of its related ad-revenue.

Here we apply our model to traces containing billions of Web requests made by millions of users to understand how information markets adopt at scale and the effect they have on total and individuals' revenue.

5.2 Data and adoption dynamics

Browsing data-sets We use anonymized HTTP traces from a university network, a neighborhood of broadband users, and a country wide mobile ISP with approximately 8k, 5k, and 3M users, respectively. We process these traces into HTTP sessions and identify publishers and aggregators for each session, for each user. Intent values $\text{Impl}_a(u)$ and $\text{Expl}_a(u)$ are computed using browsing profiles observed with partial/global views (see [29]).

Distribution of consent tracking lift . The increased value of data in the market, relative to what aggregators can infer ($r_a(u)$) is shown in Fig 6 for all user-aggregator pairs. Only 30-40% of user-aggregator pairs have $r_a(u)$ above 1.5 (threshold for a user to sell data with a benefit in a mediated market as shown in Sec. 5.1.2) and 28-34% have $r_a(u)$ above 2 (same threshold for a direct market). At first sight, the surplus appears often too small for the information market to positively benefit users. This is especially true

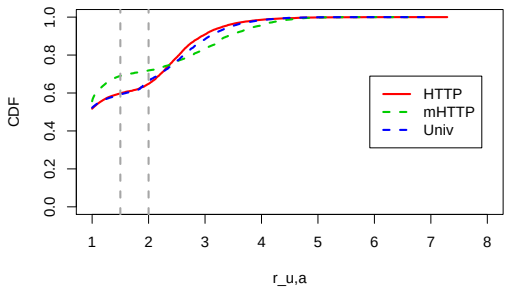

Figure 6: Distribution of consent tracking lift $r_u(a)$, defined as the ratio of advertising value with explicit intent (obtained at user consent) divided by implicit intent (obtained from tracking).

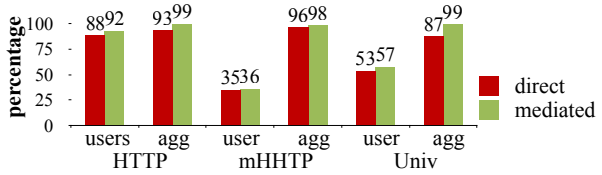

Figure 7: Percentage of users and aggregators in the market at the end of the simulation.

in the mHTTP dataset, where the aggregators appear to make accurate inferences, leading to high implicit intent which limits the value of $r_a(u)$.

Myopic best response dynamics We assume each player aims to maximize its immediate profit as it decides whether to join an information market or not: **A user** will not join an information market unless there exists an ad-network for which it can claim positive profit (*i.e.*, $r_a(u) > 2$ for a direct market, or $r_a(u) > 3/2$ for a mediated one). Note that by doing that, it also blocks tracking of other aggregators. **An ad-network** joins an information market as soon as joining the market will increase its revenue across all users and publishers in the ecosystem. Note that if many users have joined the information market (and blocked tracking) the aggregator may increase its revenue by joining the market, but still fall short of its initial revenue values before any market deployment. We analyze this in more detail in 5.4.

5.3 Spread of adoption

Fig. 7 shows the percentage of users and aggregators that eventually join the market. In contrast with what our preliminary analysis of $r_u(a)$'s distribution predicted, we see a higher adoption rate, especially among aggregators. This shows that even a small number of pairs (u, a) with high $r_u(a)$ is sufficient to generate a network effect of adoption. Indeed, more than 87% of the aggregators purchase data from at least one user, and 14-22% of those do so only because a user previously joined a market. As predicted, the mediated market has a higher percentage of adoptions with an increase of 4-8% relative to the direct market.

The same effect, although not as pronounced, is true for users: more than 35% and up to 92% see a positive profit and hence sell their data. Note that here we are only considering economic incentives for users to join, while some additional users may do so for other concerns such as privacy.

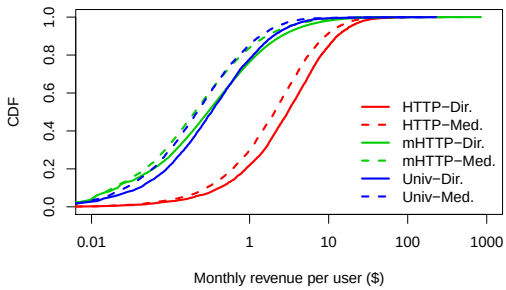

Figure 8: Monthly revenue for users that participate in the direct and mediated market.

5.4 Effects on ad-revenue

i) Markets increase overall revenue by 9-12% . Fig. 9 shows revenue at the end of the simulation normalized to initial utility, both overall and for aggregators. There is no significant difference between the direct and mediated markets in terms of overall revenue increase, with both increasing overall revenue between 9-12%.

ii) Direct markets increase revenues for users. Fig. 8 shows projected monthly revenue for users in the direct and mediated markets. Users derive more revenue in the direct market since revenue is not shared with the trusted third party. The median monthly revenue in a direct market is 50% higher than in a mediated one.

iii) User revenue is highly correlated with the number of aggregators. Monthly revenue for users in the HTTP dataset is significantly higher than in the Univ or mHTTP datasets. This stems from a high degree of correlation between aggregators the user comes in contact with and their revenue (correlation coefficients of 0.6-0.9). Indeed, the average number of aggregators per user is 43 in HTTP but only 5 in mHTTP and 9 in the Univ dataset. The higher number of aggregators per user in HTTP can be due to multiple users sharing a connection (recall this is a residential broadband trace). Next we consider how the market impacts revenue, overall and for aggregators.

iv) Aggregator revenue decreases. Revenue for aggregators decreases by 16-37% as compared to today's status-quo after information market adoption. This result contrasts with the prediction seen in Sec. 5.1 that, for a *single* transaction, aggregators always benefit from an information market. This emerges when the system is analyzed at scale because a user interacts with multiple aggregators. If it joins an information market to obtain revenue from a transaction with an aggregator (typically, one that has high $r_a(u)$) part of the consequence is that tracking is blocked and this negatively affects another aggregator revenue (typically, one with low $r_a(u)$). However, if we compare these revenues to what would happen if users, publishers, or a regulation were to block tracking with DNT [8], estimated in [29] to drop ad revenues by up to 75%, deploying an information market allows aggregators to recover from this loss. It also addresses privacy concerns of the users, as information is obtained legally and transparently.

6 RELATED WORK

The work presented in this paper lies at the intersection of online advertisements, privacy, and economics.

Privacy preserving advertisements and analytics.

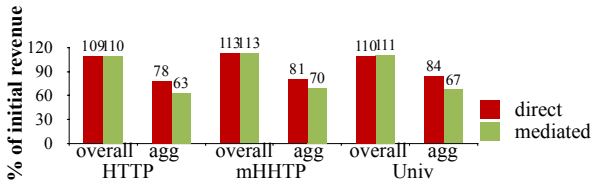

Figure 9: Total revenue normalized to initial revenue overall and for aggregators.

There has been recent interest in solutions that protect users' privacy while still enabling aggregators to provide targeted advertisements, perform analytics and personalized services [19, 23, 24, 30, 32, 44, 49]. Privad [31], Adnostic [49] and Repriv [23] are browser based systems that block access by 3rd parties to varying degrees and enable targeted ads to be shown to the user, without leaking personal information. Information markets [2, 37, 44] also address the concerns of aggregators to collect information, while allowing users to choose what information about them can get released for suitable economic compensation. Our goal in this paper is to study the viability and effectiveness of such economically driven solutions by designing, implementing and testing the solution in the wild with real users.

Privacy and economics. Our work is closely tied to work which considers personal information, through the lens of economics [22, 26, 33, 43]. Both Reznichenko *et al.* [43] and Ghosh *et al.* [26] study and design auction mechanisms; the former deals with ad-space auctions and the latter deals with direct personal information (similar to [44]). Our work is different, in that we propose and implement a system that combines privacy and performance, and study the *implications* of selling information (via a direct or mediated market) to various parties involved at a large scale using Web traces.

7 CONCLUSIONS

In this paper we investigated an information market, a system for online privacy, with focus on 3rd-party tracking, that is mutually beneficial to users and data aggregators. We discussed the requirements for an information market, and designed and implemented our system to provide privacy, Web performance, and revenue to users. We evaluate feasibility of the market via an one month field study with 63 users and observed that users actively engage with the system and are willing to sell parts of their browsing data, however these disclosures do not necessarily match up with their stated privacy concerns. Users also declared that using an information market affected their privacy attitudes.

We then proposed a model to capture the effects of information markets on advertising revenue and investigated their viability. Further, we considered the system at scale using traces containing billions of Web requests made by millions of users, to understand economic ramifications of an information market at scale and showed that it can be profitable to all parties. We found that advertising revenue increases by 9-12% overall when all players, i.e., users and data aggregators, cooperate – an observation that is at odds with most of the current beliefs about privacy preserving techniques.

REFERENCES

[1] https://www.washingtonpost.com/news/the-switch/wp/2015/10/01/heres-how-some-of-the-top-ios-9-ad-blockers-stack-up/.
[2] http://datacoup.com.
[3] http://pbb.me.
[4] http://socialdatacollective.com.
[5] http://googlesystem.blogspot.com/2009/03/behavioral-targeting-in-google-adsense.html.
[6] http://phantomjs.org.
[7] http://developer.yahoo.com/yslow/.
[8] Do not track. donottrack.us.
[9] lockerproject.org.
[10] FPDetective: dusting the web for fingerprinters. In *Computer & communications security, Proceedings of the 2013 ACM SIGSAC conference on*, pages 1129–1140. ACM, 2013.
[11] The cost of ad blocking. Technical report, Aug. 2015.
[12] TRUSTe Privacy Index 2015 Consumer Confidence Edition. *www.truste.com*, pages 1–1, Jan. 2015.
[13] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz. The Web never forgets: Persistent tracking mechanisms in the wild. *CCS '13: Proceedings of the 20th ACM conference on Computer and communications security*, 2014.
[14] A. Acquisti. Nudging privacy: T-bone economics of personal information. *IEEE S&P'09*.
[15] Adblock plus. http://adblockplus.org/.
[16] M. Ayenson, D. J. Wambach, A. Soltani, N. Good, and C. J. Hoofnagle. Flash cookies and privacy ii: Now with html5 and etag respawning. 2011.
[17] A. Aziz and R. Telang. *What is a Cookie Worth? Sixth Annual Conference on Internet Search and Innovation.*
[18] J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira. Your browsing behavior for a big mac: Economics of personal information online. In *WWW'13*.
[19] R. Chen, I. E. Akkus, and P. Francis. Splitx: High-performance private analytics. *SIGCOMM Comput. Commun. Rev.*, 43(4):315–326, Aug. 2013.
[20] D. Cvrcek, M. Kumpost, and V. Matyas. The Value of Location Information: A European-Wide Study. *Security Protocol*, 2006.
[21] Y. A. de Montjoye, E. Shmueli, S. S. Wang, and A. S. Pentland. openPDS: Protecting the Privacy of Metadata through SafeAnswers. *PLoS One*, 2014.
[22] D. S. Evans. The Online Advertising Industry: Economics, Evolution, and Privacy. *Journal of Economic Perspectives, Forthcoming*, 2009.
[23] M. Fredrikson and B. Livshits. RePriv: Re-envisioning in-browser privacy. In *IEEE S&P'11*.
[24] J. Freudiger, N. Vratonjic, and J.-P. Hubaux. Towards Privacy-Friendly Online Advertising. In *W2SP'09*.
[25] P. G. Leon et al. Why johhny can't opt out? a usability evaluation of tools to limit online behavioral advertising. Technical report, CMU, 2011.
[26] A. Ghosh and A. Roth. Selling privacy at auction. In *ACM EC'11*.
[27] A. Ghosh and A. Roth. Games and Economic Behavior. *Games and Economic Behavior*, 91(C):334–346, May 2015.
[28] P. Gill, V. Erramilli, A. Chaintreau, B. Krishnamurthy, K. Papagiannaki, and P. Rodriguez. Follow the money: understanding economics of online aggregation and advertising. *IMC '15: Proceedings of the 2015 conference on Internet measurement conference*, Oct. 2013.
[29] P. Gill, V. Erramilli, et al. Follow the money: Understanding economics of online aggregation and advertising. In *IMC'13*.
[30] A. Guha, M. Fredrikson, B. Livshits, and N. Swamy. Verified Security for Browser Extensions. *Security and Privacy (S&P), 2015 IEEE Symposium on*, pages 115–130, 2011.
[31] S. Guha, B. Cheng, and P. Francis. Privad: practical privacy in online advertising. In *NSDI'11: Proceedings of the 8th USENIX conference on Networked systems design and implementation*. USENIX Association, Mar. 2011.
[32] A. Juels. Targeted advertising ... and privacy too. In *Cryptographer's Track at RSA*, CT-RSA 2001, 2001.
[33] J. Kleinberg, C. H. Papadimitriou, and P. Raghavan. On the value of private information. *TARK'11*.
[34] B. Krishnamurthy. I know what you will do next summer. *SIGCOMM Comput. Commun. Rev.*, 40(5):65–70, Oct. 2010.
[35] B. Krishnamurthy, K. Naryshkin, and C. Wills. Privacy leakage vs. Protection measures: the growing disconnect. *W2SP'11*.
[36] K. C. Laudon. Markets and privacy. *Communications of the ACM*, 39(9):92–104, Sept. 1996.
[37] K. C. Laudon. Markets and privacy. *Commun. ACM*, 39(9):92–104, Sept. 1996.
[38] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan. Adreveal: Improving transparency into online targeted advertising. HotNets'13.
[39] V. Marotta, K. Zhang, and A. Acquisti. Who Benefits from Targeted Advertising? In *Proceedings of PrivacyCon*, Jan. 2016.
[40] F. McSherry and K. Talwar. Mechanism design via differential privacy. FOCS '07.

[41] M. Mun, S. Hao, N. Mishra, K. Shilton, J. Burke, D. Estrin, M. Hansen, and R. Govindan. Personal data vaults: a locus of control for personal data streams. *Proceedings of the 6th International COnference*, page 17, 2010.
[42] No script. http://noscript.net/.
[43] A. Reznichenko et al. Auctions in Do-Not-Track Compliant Internet Advertising. *CCS'11*.
[44] C. Riederer, V. Erramilli, A. Chaintreau, B. Krishnamurthy, and P. Rodriguez. For sale : Your data: By : You. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, HotNets-X, pages 13:1–13:6, New York, NY, USA, 2011. ACM.
[45] C. Riederer et al. For sale : Your Data By : You. *HotNets'11*.
[46] F. Roesner, T. Kohno, and D. Wetherall. Detecting and defending against third-party tracking on the web. NSDI'12.
[47] L. Shapley. A Value for n-Person Games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*. Princeton Uni. Press, 1953.
[48] L. S. Shapley. A Value of an n-Person Game. In H. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, 1953.
[49] V. Toubiana, A. Narayanan, and D. Boneh. Adnostic: Privacy preserving targeted advertising. *NDSS'10*.
[50] S. Yuan, A. Z. Abidin, M. Sloan, and J. Wang. Internet Advertising: An Interplay among Advertisers, Online Publishers, Ad Exchanges and Web Users. *arXiv.org*, June 2012.

A AUCTION MECHANISM

Here we explain the exponential auction mechanism [40] that has the truth telling properties, and has been shown to be close to optimal in terms of revenue for the seller. Note that users are independent from each other and auctions are run per user.

We denote the user as u and the set of aggregators by $\mathcal{A} = \{a_0, a_1, ..., a_m\}$. The good being sold on the market is access to the user, *i.e.,* the ability to track the user on her whitelisted sites and serve her targeted ads for a given time interval. A user's browsing behavior can be sold to multiple aggregators with no marginal cost of reproduction, hence the market can be thought of as having an unlimited supply. Intuitively, aggregator a should be willing to pay to access u as long as the price to acquire it is smaller than the additional revenue, v_a, gained by having access to u.

In the auction, we assume that each aggregator a in $\mathcal{A}$ who is interested in accessing u bids a maximum price $p_a(u)$ that it is willing to pay for access to user u. Assuming the final winning bid value is $p(u)$, every winning bidder, *i.e.,* bidders with bids higher than this value, pay $p(u)$, hence the total revenue of u is given by:

$$R\left[(p_a(u)_{a\in\mathcal{A}}, p(u)\right] = \sum_{a\in\mathcal{A}} p(u) \times \mathbb{I}_{\{p(u)\leq p_a(u)\}} \ .$$

When $p(u) > \max_{a\in\mathcal{A}} p_a(u)$, the revenue will be zero, as the price exceeds what aggregators are willing to pay. We wish to choose $p(u)$ to maximize this sum; all bids higher than this value are considered winners and hence are given access to the user. The winners pay $p(u)$.

Following [40] we first assign an initial value to $p(u)$ according to a measure $\nu(p(u))$ on $\mathbb{R}$ and then re-weigh this measure to choose the actual price used. To re-weigh, we use an exponential function that puts more weight on high value of R, according to a parameter $\varepsilon > 0$. PDF of the chosen price to track a given user is given by:

$$P(p(u)) = \frac{\exp\left(\varepsilon R\left[(p_a(u))_{a\in\mathcal{A}}, p(u)\right]\right)\nu(p)}{\int_0^\infty \exp\left(\varepsilon R\left[(p_a(u))_{a\in\mathcal{A}}, s\right]\right)\nu(s)ds}$$

A standard approach is then to choose the initial distribution of $p(u)$ according to the Lebesgue measure on $\mathbb{R}$, such that $\nu(p(u)) = 1$.

By using ε, we have added noise around the value maximizing the revenue, given the set of bids. Although it seems counter-intuitive to use a suboptimal price, [40] shows that this (1) prevents any bidder from winning more than a factor $\exp(\varepsilon)$ when cheating and (2) still reaches a revenue that is within a good bound of the optimal value if the number of aggregators is large. The expected revenue is at least $OPT - 3\frac{\ln(e+OPT\epsilon^2 m)}{\epsilon}$, where OPT denotes the optimal revenue and m is the number of buyer aggregators in the optimal case.

B VALUING USER DATA

The goal is to estimate, for a given period [0, T], the advertisement revenue that can be generated from a user based on her whitelist and impressions that she generates (frequency of visits). We want the model to be simple, and intuitively monotone in the number of impression and size of whitelist. We rely on keywords associated with websites to estimate the ad-placing strategy maximizing the total clicks generated. We assume: 1)the period is short enough so that any advertiser would like to receive a single click from the same user. Equivalently, the user would never click twice on an ad by the same advertiser in the given period. In both cases, playing an ad after is was clicked creates no revenue. 2)the period is short enough that a visit to a website at any time denotes a topic of relevance to that user for the entire period and 3)the period is long enough (or the system loses memory sufficiently quickly) so that any action (ads shown, clicks, revenue generated) from the past period is irrelevant.

We denote users by u, publisher or website by j, and advertisers by a. Each website is associated with one or several keywords that we denote by $\kappa(j)$. Similarly, each advertiser has a set of relevant keywords denoted by $\kappa(a)$. Whenever a user visits a website j and reveals that information, she effectively discloses the associated keywords. We denote by $\kappa(u) = \cup_{j\text{visited by}u}\kappa(j)$ the set of all these keywords and are all relevant to target ads to u. We assume that when u visits j and sees an ad from advertiser a, it decides independently with probability $\pi^{\text{click}}(u, j, a)$ to click on the ad, unless it has clicked on it already during that period. For simplicity we start with a constant probability, but in general this depends on the website, the advertiser (typically through the associated keywords), and the user. For instance, one may imagine that a user is more likely to click for an advertiser associated with a keyword that is associated with many sites she visits. Advertiser a can serve an ad to a user u if and only if they share one relevant keywords $\kappa(u) \cap \kappa(a) \neq \emptyset$. It typically obtains this slot through a bidding process. We will assume for simplicity that all advertisers' bids are equal to the average cost per click (CPC) for a keyword that they have in common, and we denote this value by $\text{CPC}(k)$. If this intersection has multiple keywords, then the advertiser will typically bid with the highest keywords as it denotes higher interest in the user. These values are estimated using Google AdWords keyword planner tool.

During the period, a user generates $N(u, j)$ impressions on a website j. Out of the total number of impressions created, the system decides to display a number of ads associated with a, n_a. The probability that all of these fail to generate a click is $(1-\pi^{\text{click}}(u, j, a))^{n_a}$. Hence the optimal expected value generated by clicks, depending

on which ads are played, can be found as the solution of the following optimization problem, which can be solved by a simple greedy algorithm:

$$\max \sum_{k \in \kappa(u)} \mathrm{CPC}(k) \sum_{a \text{ such that } k(a)=k} \left(1 - \pi^{\text{click}}(u, j, a)\right)^{n_a}$$
$$\text{such that } \sum_a n_a = \sum_j N(u, j) = N(u)$$

C SHAPLEY VALUE OVERVIEW

In a cooperative game, the set of players is denoted as $\mathcal{N}$. We call any subset $\mathcal{S} \subseteq N$ a *coalition* of players. For each coalition $\mathcal{S}$, we denote by $V(\mathcal{S})$ the *worth function*, which measures the total revenue produced as a result of the coalition $\mathcal{S}$.

We define the *marginal contribution* of player i to a coalition $\mathcal{S} \subseteq \mathcal{N} \backslash \{i\}$ as $\Delta_i(\mathcal{S}, V) = V(\mathcal{S} \cup \{i\}) - V(\mathcal{S})$. Note that the contribution of a player only depends on the worth function $V(\mathcal{S})$.

Shapley value determines how the total worth of the coalition, captured by $V(\mathcal{S})$, should be shared among the players in $\mathcal{S}$. More specifically, the Shapley value of player i is denoted by $\varphi_i(\mathcal{S}, V)$ and is uniquely defined by the following three axioms:

Axiom 1: (Efficiency) $\sum_{i \in \mathcal{S}} \varphi_i(\mathcal{S}, V) = V(\mathcal{S})$. This ensures that revenue assigned to the players is the total revenue created by the coalition.

Axiom 2: (Symmetry) If for all $\mathcal{S}' \subseteq \mathcal{S} \backslash \{i, j\}$, $V(\mathcal{S}' \cup \{i\}) = V(\mathcal{S}' \cup \{j\})$, then $\varphi_i(\mathcal{S}, V) = \varphi_j(\mathcal{S}, V)$. This means that two players who contribute the same amount to revenue receive an equal share of the revenue created.

Axiom 3: (Fairness/Balanced Contribution) For any $i, j \in \mathcal{S}$, j's contribution to i equals i's contribution to j, or, in other words $\varphi_i(\mathcal{S}, V) - \varphi_i(\mathcal{S} \backslash \{j\}, V) = \varphi_j(\mathcal{S}, V) - \varphi_j(\mathcal{S} \backslash \{i\}, V)$. This addresses fairness between any pair of players.

Based on the axioms above, one can show that the *Shapley value* φ can be computed as follows [47]:

$$\forall\, i \in \mathcal{S},\ \varphi_i(\mathcal{S}, V) = \frac{1}{|\mathcal{S}|!} \sum_{\pi \in \Pi} \Delta_i(S(\pi, i), V) \tag{2}$$

where Π is the set of all $|\mathcal{S}|!$ orderings of $\mathcal{S}$ and $S(\pi, i)$ is the set of players preceding i in the ordering π.

The Shapley value of a player i can thus be interpreted as the *expected* marginal contribution $\Delta_i(\mathcal{S}', V)$ where $\mathcal{S}'$ is the set of players in $\mathcal{S}$ preceding i in a uniformly distributed random ordering of $\mathcal{S}$.