Follow the Money

Understanding Economics of Online Aggregation and Advertising

Phillipa Gill
Stony Brook University
phillipa@cs.stonybrook.edy

Vijay Erramilli
Telefonica Research
vijay@tid.es

Augustin Chaintreau
Columbia University
augustin@cs.columbia.edu

Bala Krishnamurthy
AT&T Labs – Research
bala@research.att.com

Dina Papagiannaki,
Pablo Rodriguez
Telefonica Research
dina@tid.es,
pablorr@tid.es

ABSTRACT

The large-scale collection and exploitation of personal information to drive targeted online advertisements has raised privacy concerns. As a step towards understanding these concerns, we study the relationship between *how much* information is collected and *how valuable* it is for advertising. We use HTTP traces consisting of millions of users to aid our study and also present the first comparative study *between* aggregators. We develop a simple model that captures the various parameters of today's advertising revenues, whose values are estimated via the traces. Our results show that per aggregator revenue is skewed (5% accounting for 90% of revenues), while the contribution of users to advertising revenue is much less skewed (20% accounting for 80% of revenue). Google is dominant in terms of revenue and reach (presence on 80% of publishers). We also show that if all 5% of the top users in terms of revenue were to install privacy protection, with no corresponding reaction from the publishers, then the revenue can drop by 30%.

Categories and Subject Descriptors

H.1.0 [**Models and Principles**]: General

General Terms

Economics, Measurement

Keywords

Advertising, Privacy, CPM, Do-not-track, Publishers, Aggregators

IMC'13, October 23–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2504730.2504768.

1. INTRODUCTION

Online advertising plays a critical role in the Web ecosystem today. Most Web services[1] are offered for free to end users and these Web services operate by relying on revenues generated by online advertising. A lot of work has been done on different facets of online advertising: whether it be understanding the mechanisms used for advertising [25], privacy concerns [13] or combating click spam [10]. However, little is known about the *economics* of online advertising, chiefly the economics of collecting and using personal information of users for facilitating targeted advertising.

Understanding the relationship between personal information and its economic value sheds light on debates around privacy and network economics. Economics of online advertising can highlight information vectors that can lead to higher revenues, the users who supply these vectors and its impact on online privacy. Network economics literature has not considered revenues generated for Web services by the users via ads and a fine-grained characterization can lead to better models and accurate understanding of the flow of money [19].

Our key contribution is characterizing advertising revenue as a function of users' information that is used to drive online advertising, specifically how much information do different parties collect, and the value of this information to them. We begin by developing a model of online advertising revenue based on discussions with advertising professionals and literature [16, 6] (Sec. 2). Our model includes content-hosting publishers, users browsing content and aggregators tracking these users across publishers. We then perform a data-driven analysis using our model to study today's advertising ecosystem, using multiple large Web traces (Sec. 3- 4). The unique nature of our traces also lets us compare different aggregators – impossible to do by studying the clickstream of a single aggregator. To highlight the utility of understanding revenues as a function of information, we investigate what would happen to revenues if users were to adopt a privacy solution such as do-not-track (DNT), and show why such unilateral actions will not be supported by Web services (Sec. 5). Our main findings include:

Today's advertising ecosystem. In Sec. 4.2, we observe that: (a) Google is a dominant player in the online ad in-

[1]By services, we refer to online search, social networks, email as well as providing content

dustry, with presence on 80% of publishers in our datasets, with highest revenues as an aggregator but is not the top publisher in terms of revenue, and Facebook is increasing its presence around the Web with their 'Like' button, reaching 23% of publishers, (b) a few aggregators account for most of the revenue (5% accounting for 90% of revenues), however, users' contribution to advertising revenue is much less skewed (20% accounting for 80% of revenue) and (c) popular publishers account for highest revenues, while less popular ones have low revenues.

DNT/Blocking can decrease overall revenue by 75%. We learn that there can be up to 75% drop in ad revenues if blocking is near-total, without any repercussions from aggregators (Sec. 5), and just the top 5% of users (in terms of revenue contribution) blocking is enough to decrease the advertising revenue by 30%.

2. ONLINE ADVERTISING MODEL

The reason stated by online aggregators for collecting information is to increase advertising effectiveness. Hence, we focus on online advertising.

2.1 Online advertising entities

The online advertising ecosystem consists of three main entities[2]:

Users ($\mathcal{U}$) access Web content and services. We focus on content and services the user accesses for free.

Publishers ($\mathcal{P}$) host content and services that are provided free-of-charge to users. Publishers gain revenue by selling space on their Web pages to advertisers. Examples include `nytimes.com` and `slate.com`.

Aggregators ($\mathcal{A}$) map advertisers to the most effective Web page placements based on content of the Web site and any information they have about users viewing the page. To facilitate this process, aggregators track user behavior using a combination of Web-bugs, cookies, analytics *etc.*(see [18] and references therein).

Examples include DoubleClick (owned by Google) and more recently Facebook that tracks users (for social purposes) via the ubiquitous "Like" button [26]. Some large aggregators like Google also host content (*e.g.,* Youtube), hence can be publishers as well.

2.2 The role of users

In an implicit exchange for free services, users contribute to advertising revenue by viewing advertisements when they visit a publisher. We denote the number of visits a user u makes to publisher p as $\mu_u(p)$. Each visit a user makes to a Web page produces impressions that may be sold to advertisers.

2.3 Revenue for publishers and aggregators

Aggregators and publishers share advertising revenue generated by displaying ads on Web sites. We assume the aggregator retains a constant fraction of the advertising revenue (α) and passes the remaining amount on to the publisher. Google AdSense, for instance, keeps $\alpha \sim 0.32$ [21]. (We use this value in the paper.)

We consider ad revenue on a "cost-per-mille" (CPM) (*i.e.,* price for 1,000 ad impressions) basis as this is the *primary* method of purchasing targeted display ads [16]. Our model can be extended to handle cost-per-click (CPC) but this is left for future work. The amount an advertiser will pay for impressions depends on user u, ad network a and publisher p.

$$CPM(u,p,a) = RON_a \times TQM_p \times \mathcal{I}_a(u) \qquad (1)$$

Run-of-network (RON_a). RON_a is the base price for an impression in ad network a. A RON ad is a generic ad that is shown to users about whom little is known and may be shown on *any* publisher that a is affiliated with [6]. It has been observed that CPM for a *targeted* ad can go anywhere between 2-10X the price of a normal RON ad [16].

Traffic quality multiplier (TQM_p). TQM_p is a multiplier of the impression price that captures the quality of the publisher, ad location and hence the value of the impression.

User Intent $\mathcal{I}_a(u)$. The value of an impression increases as a function of the estimated purchasing intent of the user. Currently, aggregators segment users based on their interests [5], as inferred through online tracking. Certain segments are determined to have higher purchasing intent (*e.g.,* cell phone shoppers) and these users' impressions are worth more. This relative value, in turn, is reflected in the price that ad-networks charge for reaching these users via keywords and/or bid values.

We use implicit intent $II_a(u)$ to represent the intent value an aggregator can *infer* about a user. It naturally depends on the presence of an aggregator on the sites the user visits. We distinguish this from explicit intent $EI(u)$ that is computed with knowledge of *all* sites the user visits. Consider the example: user Bob visits (`espn.com, swimming.com, pets.com`). Aggregator A is present on the first two publishers, while aggregator B is present on the third one. Implicit intent for aggregator A about Bob would be limited to Bob being interested in sports, while for aggregator B, it is that Bob is interested in pets. The explicit intent $EI(u)$ is that Bob is interested in sports *and* pets. We assume that $II_a(u) \leq EI(u)$; implicit intent is capped by explicit intent. We note that while having longitudinal data helps in getting more accurate estimates of user intent, an aggregator would need to be present on *all* publishers to accurately estimate explicit intent $EI(u)$.

The total revenue[3] then of the online advertising ecosystem is the following:

$$\mathcal{R} = \sum_{u \in \mathcal{U}} \sum_{p \in \mathcal{P}} \left[\left(\sum_{a \in \mathcal{A}} \frac{\mu_u(p)}{1000} CPM(u,p,a) \right) \right] \qquad (2)$$

3. DATA ANALYSIS METHODOLOGY

We use traces[4] of HTTP traffic in multiple networks to study advertising in relation to information gathered from the users. While having access to an aggregator or a publisher's clickstream would aid our study, it would provide

[2]The actual online ad ecosystem is staggeringly complex with multiple entities (*e.g.,* data management and demand-side platforms) `http://www.liesdamnedlies.com/online_advertising_business_101.html`. We want to understand relation between information and monetization at a first order approximation and so ignore the full chain.

[3]Note that this an estimate of revenue; we use 'revenue' to refer to this estimate

[4]No personally identifiable information was gathered or used. To the extent, any data was used, it was anonymous and aggregated data.

Table 1: Summary of data sets.

Trace	Setting	Country	Users	Sessions
HTTP	Neighborhood	A (4/2011)	$\sim$ 5K	40M
mHTTP	Country	B (8/2011)	$\sim$ 3M	1.5B
Univ	Campus	C (9/2010)	$\sim$ 8K	30M

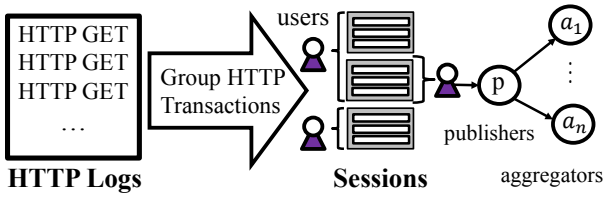

Figure 1: Data analysis pipeline to extract users, publishers and aggregators.

only a single point-of-view. In contrast, HTTP traces give us near complete visibility into the set of publishers and aggregators that the user population interacts with when they are present in the network. We also describe how we assign values to the parameters described in Sec. 2 from the data.

3.1 Data analysis overview

We use three data sets, summarized in Table 1. Both Univ and HTTP deal with traffic over a wired network, while mHTTP is traffic over a mobile network.

We first group each (anonymized) user's HTTP transactions in the HTTP traces into sessions and then identify publishers and aggregators within each session yielding a set of publishers and aggregators per user. We use the set of publishers to compute user intent ($II_a(u)$ and $EI(u)$ from Sec. 2). With intent values and values for RON_a and TQM_p, we compute $CPM(u, p, a)$ for each user-publisher-aggregator triple (Eqn. 1), and thus the overall revenue (Eqn. 2). The first two steps (Fig. 1) are described next.

3.2 Extracting HTTP sessions

We initially group HTTP transactions into sessions representing Web site visits using these heuristics.

1. StreamStructure [17]. For Univ we use the `Referer` header to group HTTP requests into sessions using the StreamStructure method proposed in [17]. This method has been shown to have precision and recall values between 0.8 and 0.9 using Alexa data [17].

2. Content-type and time. HTTP and mHTTP traces did not contain the `Referer` header so we group requests using the `Content-Type` header. We group requests between `TEXT/HTML` requests into sessions. We require that requests be more than $1s$ apart to avoid separating Web site frames into separate sessions. This method has been shown to be robust to thresholds between $0.5 - 2s$ with precision and recall values of between 0.7 and 0.8 using Alexa data [17]. We note here that while we may mis-identify a third party as a first party, the opposite would be much less likely as we use the `Content-Type`.

3. User-Agent. For mHTTP , we exploit the fact that mobile applications use HTTP for their communications and set the `User-Agent` field to indicate which application is making the request. When the `User-Agent` is not a mobile browser we group requests for the same application based on `User-Agent`.

For the three methods above we also exclude known third party domains (*e.g.,* those identified in [18]) from consideration as a publisher. We also require that sessions contain more than one request as most Web pages today contain more than one object (*e.g.,* images).

Identifying publishers and aggregators. After sessionizing HTTP transactions, we assign the first domain in the session to be the publisher. We consider any domain hosted on a different AS than the publisher to be a third party. We use RIPE's `whois` to perform the IP to AS mapping.

Using AS numbers to distinguish publisher and aggregator organizations has limitations. For example, CDNs hosting embedded content are classified as aggregators. Indeed, we observed CDNs as some of the most common aggregators. That emphasizes their potential to enter the aggregation business [4]. To mitigate the impact of CDNs on our results we exclude well known CDNs from consideration as aggregators (*e.g.,* Amazon (AS16509/14618), Akamai (AS16625/20940), and Limelight(AS22822)). Secondly, publishers and aggregators that are owned by one entity (*e.g.,* Gmail and Doubleclick) cannot be disambiguated. This is not a problem as we assume they share information (Google has a unified privacy policy [23]). We attempted using an alternate technique, combining ADNS entries and AS numbers. However, it divided some popular organizations (*e.g.,* Microsoft uses both msecd and nsatc as ADNS domains). For cases where multiple publishers are hosted on the same AS (*e.g.,* a CDN) we identify publishers using their host domain.

3.3 Computing intent $\mathcal{I}_a(u)$

The ability to profile users based on their purchasing intent is what drives advertisers to pay more for targeted advertisements. Advertisements targeted to specific audience segments can command 2-10X the base RON_a price (depending on the predicted purchasing intent of the segment) with an average increase of 3.3X [16]. Reverse engineering the data mining algorithms used to determine purchasing intent is beyond the scope of this paper. However, to lend realism to our calculations, we rely on values assigned by ad-networks themselves to key-words in combination with categories that can be assigned to publishers. So we use the following process.

1. Categorize user-visited publishers Users' interests and intent can be inferred based on publishers visited (*e.g.,* `espn.com`: sports) in the same way as aggregators do it [12]. Using Alexa's standard site categorization of our datasets mimics advertisers planning campaigns. When publishers appear in multiple categories in Alexa (*e.g.,* `bbc.co.uk`, bbc.co.uk: arts and news) we pick the highest ranking category. Note that this is conservative as publishers who map to different categories can lead to higher revenue.

2. Determine intent values for categories. Once we know the Web sites categories a user visits, we need to convert these categories into a multiplier value between 2 and 10 (as this is the increase in value of an impression due to targeting). We use *suggested* bid amounts for the categories provided by the Google AdWords Contextual Advertising Tool[5], to estimate the *relative value* of different categories.

[5]Can be found at `adwords.google.com`.

Once we have the bid amount for each category, we normalize by the highest category value and rescale into the range 2-10. Hence, every publisher gets mapped to a category and using the category, an intent value is assigned.

3. Compute intent. We use the set of publishers a user visits to calculate two different intent values for the user: *implicit* $II_a(u)$ for an aggregator a, computed by taking the *average* of intent value of publishers that u visits where a is present as a third party (results in an average $II_a(u)$ of 3.1 in the Univ and HTTP datasets and 3.8 in mHTTP) and *explicit* $EI(u)$ to be max of the average intent value across *all* publishers u visits or $II_a(u)$. The difference between the intents approximates the added value of having visibility into the full set of sites visited. Relative frequency of visits to different publishers can also be part of our calculations.

3.4 Additional parameters

Traffic quality multiplier (TQM_p). TQM_p captures additional factors (ad placement, quality of publisher) impacting the value of impressions. Capturing all of these factor is beyond the scope of this work. We instead, focus on TQM_p as based on the quality of the publisher. Reasonable values of TQM_p are 2 for popular publishers (*e.g.,* New York Times) or 0.1 for disreputable sites (*e.g.,* illegal file hosting). Thus, we assign publishers appearing in the top 500 sites on Alexa $TQM_p = 2$. We assign publishers with IPs on a DNS blacklist[6] $TQM_p = 0.1$. We assign the remaining publishers $TQM_p = 1$. We note that a publisher may be outside the Alexa top 500 but can still be a prime/reputable site, but the penalty for mis-classifying such sites is a factor of two. However, we want to ensure we never assign disreputable sites (as given by the blacklist) a high score, hence we assign a number that differs by an order of magnitude (0.1)

Run-of-network (RON_a). We use a value of $1.98 for RON_a as this is the average run-of-network price found in advertising literature [6].

Limitations: We used published numbers for our parameters but they may vary in practice; our results are *not* meant to predict absolute values. Hence, we focus on distributions and trends in ad revenues that are not impacted by scalar values such as TQM_p or RON_a. There may be issues with relying on a certain method to classify publishers/aggregators or on certain resources (Alexa, Adwords) but methodology remains the same.

4. TODAY'S ADVERTISING REVENUE

We now combine the advertising model (Sec. 2) with the datasets (Sec 3) to see: (1) how much do aggregators know about users through tracking? (2) which users, publishers and aggregators generate the most ad revenue? We mostly present results from mHTTP due to space constraints.

4.1 How much do aggregators know?

Aggregators are able to estimate intent accurately for 50% of users. Fig. 2 shows the ratio of explicit to implicit intent for user-aggregator pairs. Recall, that for each user, the aggregator infers intent based on the subset of sites where the aggregator is present as a third party. We find that more than half of aggregators in all datasets

[6] `http://dnsbl.inps.de`

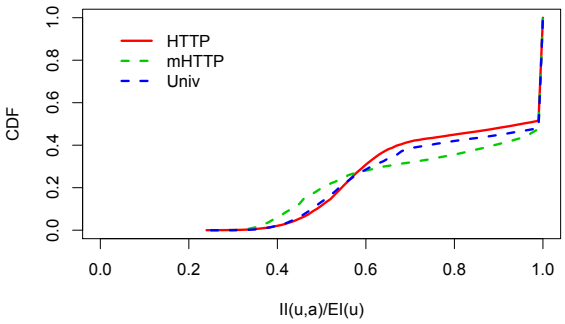

Figure 2: CDF of inferred intent ($II_a(u)$) normalized by explicit intent ($EI(u)$).

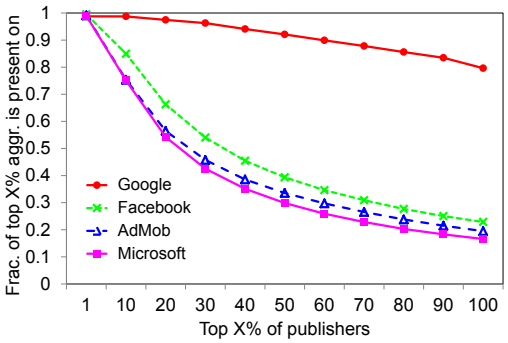

Figure 3: Presence of the top four aggregators on publishers with the most revenue, mHTTP.

inferring the correct value of $EI(u)$ for those users. This accuracy stems from many users visiting sites in a small number of categories (median: 2.2 categories, mHTTP).

Aggregators know most about popular sites. In Table 2, we show the reach of top (in terms of revenue) aggregators across all publishers in our datasets. Maintaining presence on many publishers requires aggregators to build and maintain business relationships. Fig. 3 shows the fraction of publishers the top four aggregators are present on for varying numbers of top (in terms of popularity) publishers. Top aggregators are focusing on popular publishers with the top aggregators present on more than 70% of the top 10% of sites. As we consider less popular sites presence by top aggregators correspondingly decreases with Facebook dropping from presence on 85% of the top 10% of publishers to presence on only 23% of all publishers. This suggests that studies considering third party tracking on popular publishers (*e.g.,* [18]) are seeing an upper bound on tracking. In terms of implications to privacy, we find most aggregators are present on a low number of publishers (Google being an exception, Table 2).

4.2 How valuable is this information?

Ad revenue is generated by many users. Ad revenue generated by users is only slightly skewed, with 90% of revenue derived from 55% of HTTP and 35% of mHTTP and Univ users, respectively (Fig. 4). We find strong correlation between user revenue and the number of sessions per user with a correlation (r-value) of 0.64 for mHTTP. Unsurprisingly, users who browse more are more valuable in the impression-based revenue model.

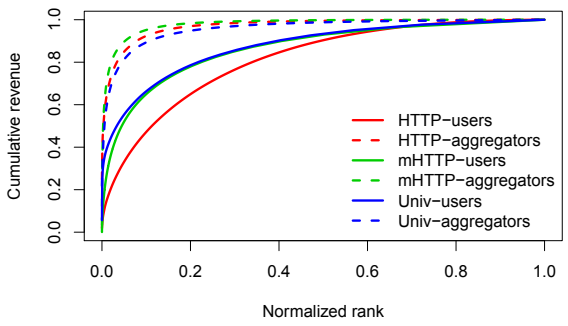

Figure 4: Cumulative fraction of revenue attributed to each aggregator and user.

Table 2: High revenue aggregators(mHTTP).

Aggregator	Frac. Rev.	Frac. Users	Frac. Pubs.
Google	0.18	0.17	0.80
Facebook	0.06	0.09	0.23
GlobalCrossing (AdMob)	0.04	0.11	0.19
AOL	0.03	0.04	0.07
Microsoft	0.03	0.04	0.17
Omniture	0.03	0.05	0.07
Yahoo! (AS42173)	0.03	0.04	0.07

Which information vectors are most lucrative? As mentioned earlier, number of sessions play a more central role in determining revenue. However, we find that some categories are more lucrative than others ('Recreation' for mHTTP), but this is reflected in the bid prices for different categories (Sec. 3.3). In addition, the bid prices also reflect the relative popularity of the categories (publishers classified as 'Recreation' are the most visited in mHTTP). Simultaneously, we find that publishers that are least visited belong to categories that have the lowest bid prices. From a privacy perspective, if we consider k-anonymity, then the most popular publishers (and most lucrative) would be the most private while the least lucrative/popular would be less private.

Most popular publishers do not necessarily generate most revenue. Table 3 shows the top publishers in the mHTTP dataset. We find that while Google (`google.com`) is the most visited publisher with 18% of users visiting Google as a publisher[7], Facebook (`facebook.com`) actually generates the most revenue: 9%. We see Facebook's CDN `fbcdn.net` also generating significant revenue since it also serves Facebook Web pages. Revenue is correlated with the number of aggregators present on each publisher, in the mHTTP dataset, we find a correlation of 0.61 (r-value) between number of aggregators and revenue per publisher.

Google is the top aggregator Table 2 shows the top aggregators in the mHTTP dataset. As in previous work [18], we observe Google playing an active role as an aggregator. Google is present on significantly more publishers than the other aggregators, with presence on 80% of publishers in the mHTTP dataset. Fig. 4 shows that advertising revenue is concentrated by a few aggregators with the top 5-10% of aggregators getting 90% of the ad revenue. Facebook also ranks highly as an aggregator reaching 9% of users with presence on 23% of first parties in the mHTTP dataset.

[7]Domain is used to identify publishers (google.co.uk/google.com are different) and given possible overlap among users we cannot sum the fraction.

Table 3: High revenue publishers (mHTTP).

Publisher	Frac. Rev.	Frac. Users	Category
facebook.com	0.09	0.15	society
google.co.uk	0.04	0.11	computers
bbc.co.uk	0.03	0.07	arts
fbcdn.net	0.03	0.13	society
twitter.com	0.03	0.04	computers
yahoo.com	0.03	0.04	computers
google.com	0.02	0.18	computers

5. REVENUE WITH PRIVACY

We study how unilateral privacy preserving actions by users affect advertising revenues. We consider blocking technologies, that disrupt tracking by third parties, downloading of online ads [3], deny cookies, limit Javascript execution to selective sites [22] as well as obfuscation methods where the key idea is to either inject noise in services that profile users, like search [15] or mobile apps [14] or to impersonate users [1]. The users' privacy is protected as their 'behavior' is obfuscated/attributed to someone else. The aggregators clearly lose out as the data they obtain is corrupted.

We note that these methods may not be effective and few might use them [11, 9]. Likewise, measures like the Do Not Track initiative [2] relies on aggregators honoring the intent. In addition, publishers and aggregators may retaliate by refusing service when they detect blocking, decreasing the utility for the user. Whatever the outcome might be in such a cat-and-mouse game, the intention from the users' is clear – to opt-out of tracking. Hence blocking prevents aggregators from tracking users and reduces revenue generated by targeted advertisements. We quantify this loss.

5.1 Quantifying revenue loss

The amount of information available to aggregators may differ based on privacy protection measures used and can be captured in the model from Sec. 2.

The cost of blocking When users deploy privacy preserving measures they may experience decreased functionality. Modeling the utility decrease due to blocking is beyond the scope of this paper.

User Intent $\mathcal{I}_a(u)$. Recall that user intent captures the interests of the user (Sec. 2). Privacy preservation techniques hinder aggregators from being able to infer the intent of users through tracking. This gives two potential values for $\mathcal{I}_a(u)$ in our model:

$$\mathcal{I}_a(u) = \begin{cases} II_a(u) & u \text{ and } p \text{ do nothing} \\ 1 & \text{otherwise} \end{cases} \quad (3)$$

Implicit intent ($II_a(u)$) is as described in Sec. 3.3. Recall that implicit intent ($II_a(u)$) is what aggregators can infer, while explicit intent ($EI(u)$) can consist of all the information the user possesses. And when the user or publisher block tracking there is no increase in CPM as a result of intent, hence it is set to 1.

Quantifying the cost of blocking Fig. 5 shows how much value is currently derived from implicit intent which stands to be lost if users block. The average value of $II_a(u)$ is 4.2

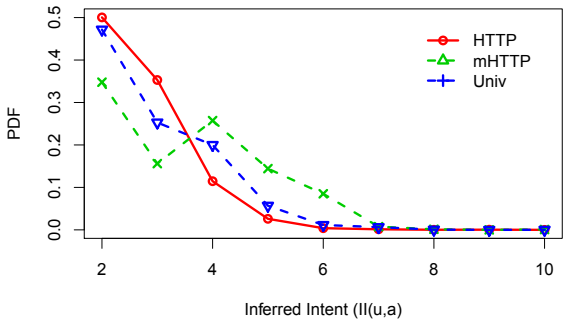

Figure 5: Distribution of implicit intent, $II_a(u)$.

in the HTTP, 3.8 in mHTTP and 3.1 in the Univ traces, respectively. Indeed, when we compute revenue with all users blocking (*i.e.,* $\mathcal{I}_a(u) = 1$) revenue decreases by a factor of 4.2 in the HTTP, 3.8 in mHTTP, and 3.2 in the Univ traces, respectively. A large population of users blocking – in the worst case, if the Do Not Track (DNT) header [2] became default – would represent a significant threat to advertising revenue. If proposals like DNT are honored by aggregators this may lead to lowered quality of service as the publisher will lose out on additional revenues. Blocking also poses the potential to decrease functionality of Web sites for users (*e.g.,* blocking Javascript via NoScript [22]). Hence, for these reasons, it can be argued that most users will not take the extreme step of blocking entirely. However, we find that even if 5% of the top users (Fig. 4) block, the revenue drop is between 35%-60%. Regarding obfuscation, assuming that incorrect targeting is worse than no targeting, the drop in revenues due to blocking will be a lower bound on *revenue loss* due to obfuscation.

6. RELATED WORK

Much work has been done on understanding the effectiveness of behavioral targeting [6, 29], users' attitudes towards targeting [20], and designing systems for mining interests of users for targeting [8]. The actual behavioral targeting mechanisms are not known as they tend to be proprietary. Likewise, recent work has focused on the problem of combating clickspam [10] and characterizing mobile advertising [28]. In addition, past work has also focused on understanding the role online ads play in decision making [24, 27] or how to properly set CPM values [7]. Our paper tries to shed light on part of the advertising landscape using real data from an economics perspective – understanding advertising revenue as a function of user's information.

7. CONCLUSIONS

Using HTTP traces, we looked into the relationship between how much information is collected by aggregators and quantified how valuable this information can be. We developed a simple revenue model for targeted advertising to aid us. Our results indicate that revenue is highly skewed towards few aggregators, while it is not that skewed from a user perspective. We also studied possible revenue drop if users were to adopt unilateral privacy protection.

Acknowledgments

We would like to thank the reviewers and our shepherd, Craig Partridge for helpful comments. We would like to thank Martin Arlitt for providing us with the university Web traces and infrastructure for analysis. We also thank Glenn Ellison, Jonathan Smith, John Byers and Georgios Siganos for comments on a draft of this work. Phillipa Gill was supported by the AT&T VURI program.

8. REFERENCES

[1] Bug me not. `bugmenot.com`.
[2] Do not track. `donottrack.us`.
[3] Adblock plus. `http://adblockplus.org/`.
[4] Akamai (Press Release). Akamai introduces advertising decision solutions; Announces agreement to acquire Acerno, 2012. `http://www.akamai.com/html/about/press/releases/2008/press_102108.html`.
[5] AudienceScience. The audiencescience targeting segments, 2012. `www.audiencetargeting.com`.
[6] H. Beales. The value of behavioral targeting. *National Advertising Initiative*, 2010.
[7] H. Cannon. Addressing new media with conventional media planning. *Journal of Interactive Advertising*, Vol 1, No 2, 2001.
[8] Y. Chen, D. Pavlov, and J. F. Canny. Large-scale behavioral targeting. In *In Proc. of ACM SIGKDD*, KDD '09, 2009.
[9] Cybernet News. Less than 3% of Firefox users block Ads, 2008. `http://cybernetnews.com/less-than-3-of-firefox-users-block-ads/`.
[10] V. Dave, S. Guha, and Y. Zhang. Measuring and Fingerprinting Click-Spam in Ad Networks. In *Proceedings of the Special Interest Group on Data Communication (SIGCOMM)*, Helsinki, Finland, Aug 2012.
[11] 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.
[12] Google. Your data on Google: Advertising, 2012. `http://www.google.com/goodtoknow/data-on-google/advertising/`.
[13] S. Guha, B. Cheng, and P. Francis. Privad: Practical Privacy in Online Advertising. *Proc. of NSDI*, 2011.
[14] P. Hornyack et al. "These Aren't the Droids You're Looking For": Retrofitting Android to Protect Data from Imperious Applications. In *Proc. of ACM CCS*, 2011.
[15] D. Howe and H. Nissenbaum. Trackmenot: Resisting surveillance in web search. In I. Kerr, C. Lucock, and V. Steeves, editors, *On the Identity Trail: Privacy, Anonymity and Identity in a Networked Society*. 2008.
[16] A. Hunter, M. Jacobsen, R. Talens, and T. Winders. When money moves to digital, where should it go? identifying the right media-placement strategies for digital display. Technical report, Comscore and ValueClick Media, 2010.
[17] S. Ihm and V. S. Pai. Towards understanding modern web traffic. In *In Proc. of the ACM SIGCOMM IMC*, 2011.
[18] B. Krishnamurthy and C. Wills. Privacy Diffusion on the Web: A Longitudinal Perspective. In *Proc. ACM WWW*, page 541. ACM Press, 2009.
[19] R. T. B. Ma, D. M. Chiu, J. C. S. Lui, V. Misra, and D. Rubenstein. On cooperative settlement between content, transit, and eyeball internet service providers. *IEEE/ACM Trans. Netw.*, 19(3):802–815, June 2011.
[20] A. M. McDonald and L. F. Cranor. Americans' attitudes about internet behavioral advertising practices. In *WPES*, 2010.

[21] N. Mohan. The AdSense revenue share, 2010. http://adsense.blogspot.com/2010/05/adsense-revenue-share.html.

[22] No script. http://noscript.net/.

[23] NYTimes. Europe presses google to change privacy policy, 2013. http://www.nytimes.com/2012/10/17/business/global/17iht-google17.html?pagewanted=all&_r=1&.

[24] P. Patwardhan and J. Ramaprasad. Rational integrative model of online consumer decision making. *Journal of Interactive Advertising*, Vol 6, No 1, 2005.

[25] A. Reznichenko et al. Auctions in Do-Not-Track Compliant Internet Advertising. *ACM CCS*, 2011.

[26] R. Richmond. As ŚlikeŠ buttons spread, so do facebookŠs tentacles, 2011. http://bits.blogs.nytimes.com/2011/09/27/as-like-buttons-spread-so-do-facebooks-tentacles/.

[27] S. Rodgers and E. Thorson. The interactive advertising model: How users perceive and process online ads. *Journal of Interactive Advertising*, Vol 1, No 1, 2000.

[28] N. Vallina-Rodriguez, J. Shah, A. Finamore, Y. Grunenberger, K. Papagiannaki, H. Haddadi, and J. Crowcroft. Breaking for commercials: characterizing mobile advertising. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, IMC '12.

[29] Y. Yan et al. How much can behavioral targeting help online advertising? In *In Proc. of WWW*, 2009.