Do-Not-Track and the Economics of Third-Party Advertising*

Ceren Budak
Microsoft Research

Sharad Goel
Stanford University

Justin M. Rao
Microsoft Research

Georgios Zervas
Boston University

Abstract

Retailers regularly target users with online ads based on their web browsing activity, benefiting both the retailers, who can better reach potential customers, and content providers, who can increase ad revenue by displaying more effective ads. The effectiveness of such ads relies on third-party brokers that maintain detailed user information, prompting legislation such as do-not-track that would limit or ban the practice. We gauge the economic costs of such privacy policies by analyzing the anonymized web browsing histories of 14 million individuals. We find that only 3% of retail sessions are currently initiated by ads capable of incorporating third-party information, a number that holds across market segments, for online-only retailers, and under permissive click-attribution assumptions. Third-party capable advertising is shown by 12% of content providers, accounting for 32% of their page views; this reliance is concentrated in online publishing (*e.g.*, news outlets) where the rate is 91%. We estimate that most of the top 10,000 content providers could generate comparable revenue by switching to a "freemium" model, in which loyal site visitors are charged $2 (or less) per month. We conclude that do-not-track legislation would impact, but not fundamentally fracture, the Internet economy.

Keywords: *e-commerce, privacy, competition, advertising*
JEL Codes: *L10, M37, C93*

*Any views and opinions expressed herein are solely those of the authors and do not reflect those of the Microsoft Corporation, Stanford University, or Boston University. We thank Susan Athey, Avi Goldfarb, David Pennock, David Rothschild, Matthew Salganik and Catherine Tucker for helpful comments and discussion.

1 Introduction

In recent years it has become not only technologically possible but also commonplace to tailor online advertisements to a user's interests and past browsing behavior, a practice known as behavioral targeting. To implement such targeting, various ad networks, ad exchanges, and other so-called "third parties" that are not directly associated with the website on which an ad is shown track users as they browse the web (Mayer and Mitchell, 2012). This information is then made available to advertisers in real-time ad auctions, who subsequently bid to show ads to users based on their compiled online profile (Google, 2011). A bookseller, for example, may bid to show an ad to a user on a sports site after learning that she had recently visited an online magazine aligned with the bookseller's target audience. Relative to simply displaying the same ad to all users of a site, behaviorally targeted ads have much higher response rates, thus benefiting both advertisers who can more efficiently reach consumers, as well as content providers who can better monetize their traffic (Yan et al., 2009; Goldfarb and Tucker, 2011a; Farahat and Bailey, 2012; Johnson, 2013). The demand from both advertisers and content providers for such highly personalized advertising has led to over 90% of the top 500 websites sharing information with third-party trackers.

Though the practice has been widely adopted by advertisers and content providers, third-party tracking can potentially degrade user privacy. For example, queries on a health website could trigger ads from a pharmacy in subsequent web browsing sessions, leaking sensitive information to a firm with which the user has never directly interacted. Hundreds of entities provide tracking services, and consumer data are bought and sold in loosely regulated marketplaces (Roesner et al., 2012). The detailed nature of the collected information makes it difficult, if not impossible, to anonymize—researchers have shown that browsing histories can often be linked to names, addresses and other personally identifying information (Krishnamurthy et al., 2011; Reisman et al., 2014). Additionally, web sites at times inadvertently provide trackers with information that they intended to keep private, such as usernames and email addresses, because the information is embedded in unencrypted URLs.[1] The collection, sharing, and selling of data has also been shown to facilitate identity fraud and spam (Mayer and Mitchell, 2012). In addition, the companies that track users' movement across almost all of the popular Web sites (Krishnamurthy and Wills, 2009) form a small-world network (Gomer et al., 2013). While this characteristic allows efficiency in spreading user information and delivering targeted ads, it also introduces further privacy and security concerns.

Reacting to such concerns, consumer and privacy advocates have proposed legislation, known broadly as "do-not-track," that would limit or ban third-party tracking. Citing the benefits of third-party data for ad effectiveness, advertisers and content providers have largely opposed such bans, arguing that do-not-track would adversely impact their businesses and, in turn, harm consumers. Such economic concerns are indeed grounded in recent empirical findings. For example, Goldfarb and Tucker (2011b) use survey data to evaluate the impact of the 2002 European Union "Privacy and Electronic Communications Directive." They find that on average, stated purchasing intent declines 65% in the E.U. compared to control countries. Similarly, Johnson (2013) uses auction logs from a real-time display advertising exchange to counterfactually simulate the impact of privacy policies on ad prices, and finds that a full restriction would reduce prices by about 40%.[2]

While past work has focused on the direct effects of privacy policies on third-party advertising, here we examine the effects of such policies on the Internet economy as a whole, both from the perspective of advertisers and content providers. Our primary methodological approach is to analyze the web browsing histories of 13.6 million users for the 12 months between June 1, 2013 and May 31, 2014. To study the advertiser side of the equation, we

[1] For the same reason, searches on sites are often observable to trackers and linkable to cookies and usernames.

[2] As Johnson notes, however, the actual restrictions placed by various incarnations of do-not-track vary widely, with his estimated impact ranging from 4–40%.

first identify in our data 321 million shopping sessions (i.e., visits to the 10,000 most popular commerce sites). For each shopping session, we then determine the primary driver of the consumer to the retail site. We find that the vast majority of shopping sessions begin with web searches, search ads, email marketing, or direct navigation to the site, none of which rely on third-party data. In fact, perhaps surprisingly, display ads, advertising that could potentially use third-party information, accounts for only 3% of shopping sessions. Moreover, only 7% of the retailers we study receive more than 10% of their traffic from such third-party capable ads. Next, we consider the possibility that certain segments of the e-commerce market are particularly reliant on third-party capable advertising. We use topic modeling (Blei et al., 2003) to algorithmically cluster retailers into 54 segments (*e.g.*, sporting goods, home improvement, and books), and find that no market segment gets more than 7% of traffic from third-party capable advertising, with most close to the group mean. Finally, we repeat our analysis on the 45% of businesses in our sample that have a physical location. For these offline retailers, we find the share of sessions attributed to third-party capable advertising increases somewhat, to 3.9%.

These results are based on a last-click attribution model, in which we associate each shopping session with the most recent event in a user's browsing history that preceded it. With such an inference scheme, however, there is a worry that seemingly direct navigation could have in fact resulted from past ad exposure. Such misattribution could be particularly large for ad campaigns that build brand awareness, particularly for firms that do business offline (Lewis et al., 2014). To address these concerns, we perform two robustness checks. First we consider the following, more conservative click-attribution model: we take all direct navigation sessions and look back one month into a user's browsing history; if, during that 28-day window the user clicked on an ad for the retailer, we credit the shopping session to the ad. We find that this increases the percentage of shopping sessions attributed to third-party capable advertising from 3% to 3.4%. While this 13% increase is surely of interest to firms assessing their advertising effectiveness, it is a relatively small difference in the context of overall traffic.

Turning to content providers, we consider the ten million domains visited by users in our sample. Of these sites, 12% regularly show third-party capable advertising. Such sites, however, are disproportionately popular, and account for 32% of aggregate traffic. In contrast to our analysis of retailers, certain segments of content providers—particularly, online publishers, such as Yahoo and the Huffington Post—are substantially more likely than average to show third-party advertising. Specifically, 48% of online publishers, accounting for 81% of all online publisher traffic, show third-party capable ads.

Given the conventional wisdom that the Internet is by-and-large ad-supported (Deighton and Quelch, 2009), and given the benefits of third-party tracking for ad effectiveness, how is it that two-thirds of Internet traffic comes from sites that do not show third-party ads? To explain this apparent incongruence, we note that many of the largest web sites either target ads based on information that users explicitly provide to the site, as in the case of Google and Facebook, or have alternative monetization models, as in the case of Craigslist and Wikipedia.

To better understand the extent to which content providers that show third-party advertising would be adversely affected by policies that limit the effectiveness of such ads, we consider their ability to generate revenue via alternative sources. Specifically, we consider one of the most prevalent alternatives to advertising in the marketplace today: a metered paywall ("freemium") model, in which site visitors only pay subscription fees to consume content in excess of a modest daily or monthly allotment. In fact, many of the largest online news outlets (*e.g.*, the New York Times and the Wall Street Journal) have already adopted this model. A necessary condition to successfully adopt a freemium model is a set of loyal users who regularly visit the site, and who consequently would be charged subscription fees. Among the top 10,000 sites that show third-party capable advertising, we find that typically

15% of users visit the site at least 10 times per month. If one-fourth of such loyal users ultimately subscribe to the site, we estimate that a monthly fee of $2 would generate revenue comparable to that earned from third-party capable advertising, based on current ad rates. Moreover, we find that consumers typically only regularly visit 2–3 sites that feature third-party capable advertising, limiting how much consumers would need to pay if such sites switched to a freemium model. However, we also find that outside of these top 10,000 sites, only a small percentage of users are loyal, suggesting that a freemium model is not feasible for all content providers.

Our empirical analysis focuses on only one-half of the do-not-track cost-benefit equation, assessing the potential economic benefits of third-party tracking. Moreover, our analysis is based on current market conditions, as it is difficult to predict market responses to industry realignment and technological advances (*e.g.*, micropayments, or changes in tracking technology). Additionally, there is not simply a single, well-defined do-not-track policy, but rather a variety of proposals that fall under this umbrella name.[3] For these reasons, we cannot offer firm guidance on whether privacy legislation such as do-not-track should be enacted. Nonetheless, our results suggest that the benefits of third-party information to retailers and content providers, while certainly not absent, are smaller than generally believed. Our work thus constitutes a novel point of reference for future studies of privacy legislation. Furthermore, by conducting one of the most comprehensive studies of what drives online shopping, our results provide a basis for ongoing marketing and advertising research.

[3] These proposals vary on quite consequential factors, such as opt-in vs. opt-out tracking agreements, and the precise definition of what constitutes a "third party." For example, if a user is browsing the New York Times, but logged into their Gmail account, would Google's ad-exchange tracker be considered a third or first party?

2 Data and Methods

Our primary analysis is based on web browsing records collected via the Bing Toolbar, a popular add-on application for the Internet Explorer web browser. In 2013, Internet Explorer was the second most popular browser in the United States, with the independent analytics firm StatCounter estimating that the browser accounted for 25% of U.S. pageviews.[4] Upon installing the toolbar, users can consent to sharing their data via an opt-out agreement, and to protect privacy, all shared records are anonymized prior to being saved on our system. Each toolbar installation is assigned a unique identifier, giving the data a panel structure. While it is certainly possible that multiple members of a household share the same browser, we follow the literature by referring to each toolbar installation as an "individual" or "user" (Gentzkow and Shapiro, 2011; De los Santos et al., 2012).

As with nearly all observational studies of individual-level web browsing behavior, our study is restricted to individuals who voluntarily share their data, which likely creates selection issues. These users, for example, are presumably less likely to be concerned about privacy. Moreover, though our panelists did not report any demographic information, it is generally believed that Internet Explorer users are on average older than the Internet population at large. Instead of attempting to re-balance our sample using difficult-to-estimate and potentially incorrect weights, we acknowledge these shortcomings and note where they might be a concern.

2.1 Data description

Our data contain detailed information on the web browsing activity of 13,560,257 U.S.-located users over a one-year period, from June 1, 2013 to May 31, 2014. Each webpage visit generates a record containing the URL of the requested page (*e.g.*, `http://www.amazon.com`),

[4] This estimate is based on visits to three million webpages that StatCounter tracks. For more on the methodology, see http://gs.statcounter.com/faq#methodology.

an anonymized id for the user viewing the page, the time at which the page was requested, and a unique identifier for the browser window or tab in which the page was rendered. Additionally, if the pageview was initiated by an HTTP redirect, the initial URL that caused the browser to display the page is logged. This information is particularly useful for detecting ad clicks, as redirects are commonly used in display advertising to deliver and track ads (by both the hosting domain and third parties). For example, when a consumer clicks on a display ad, instead of being directly sent to the advertiser's website, an HTTP request is typically first made to the web server of the party responsible for delivering the ad (*e.g.*, DoubleClick); subsequently, and almost transparently to the user, the party serving the ad records the ad click, and then redirects the user's browser to the advertiser's web site.[5] Finally, each pageview record contains all HTTP requests initiated by the page to load additional assets (*e.g.*, images and stylesheets) that are needed to render it. As with the redirects, these asset requests help us determine the presence of advertising; in particular, assets originating from known ad servers indicate the presence of one or more display ads on the page.

2.2 Classifying shopping sessions

Starting with the raw browsing data, we use the Open Directory Project (ODP, dmoz.org) to help identify retail shopping sessions. The ODP is a collective of tens of thousands of editors who hand label websites into a classification hierarchy, 45,000 of which are classified under "shopping". We focus on the 10,000 most popular such shopping sites, which in aggregate account for over 99% of traffic to the full set of 45,000. When a user visits any one of these top 10,000 retailers, we call that visit, along with all subsequent, uninterrupted visits on the same domain, a single shopping session. Though we do not know whether any financial transaction ultimately occurred, a shopping session at the very least indicates an important first step in the purchase process. In total we identify 320,889,786 shopping sessions in our

[5] For more on HTTP redirects, see `http://www.w3.org/Protocols/rfc2616/rfc2616.html`.

sample.

For each such shopping session, we classify it into one of eight categories based on the means through which the user arrived at the site: direct navigation, organic search, search advertising, email marketing, social advertising, display advertising, coupon (or "deal finder") site and organic link referral. Our classification strategy considers the referrer URL associated with each shopping session, various features of the first URL in a session, and the redirect URL (if any) that initiated the session. Though we only briefly describe this classification process below, we note that it is both labor intensive and technically challenging, as a myriad of pattern-matching rules must be developed to handle each case.

We categorize as *direct navigation* instances where the URL for the retail site is directly entered into the browser's location bar, or the user reaches the site via a bookmark, both of which are identified by the absence of a referrer URL. We also classify web searches for specific retailer names, often referred to as *navigational searches* (Broder, 2002), as direct navigation, since it indicates the user is seeking out a single retailer based on prior knowledge of the retailer's name.[6] Sessions that are initiated via web searches are identified by matching the referrer URL against a list of search engines. Moreover, we can accurately distinguish between sponsored (paid) and organic (non-paid) search by using distinctive features of the referrer and redirect URLs. *Email ads* are image or text links embedded in the content of promotional email messages (*e.g.*, an email with a Groupon deal), and we similarly detect them by matching the referrer URL to a list of known email providers and examining the redirect URL for telltale signs of such advertising.[7] We categorize shopping sessions originating from social networks—Facebook being the dominant example—as driven

[6] The search query is typically present in the referrer URL, which allows us to identify navigational searches. We would miss navigational searches using the nickname of the site that does not appear in the web address.

[7] Retailers also send their customers receipts, shipping updates, and other non-marketing information via email. We exclude sessions originating from such email messages from our analysis as they are unlikely to lead to new purchases.

by *social ads*. To detect *display ads*—graphical ads typically paired with textual content—we match the redirect URL to a comprehensive list of ad servers maintained and updated weekly by AdBlock Plus, a popular open source browser extension to block such advertising. Online retailers receive a small, but significant, number of clicks from sites that distribute digital coupons (*e.g.*, `http://www.retailmenot.com`), and we classify these shopping sessions as initiated by *coupon site referrals*. Finally, *organic link referrals* are non-paid, site-to-site links (*e.g.*, from PayPal to eBay), and are identified by cross-site traversals that do not trigger any of our ad-detection rules, such as going through a known ad-server.

2.3 Constructing retail segments

Much of our analysis occurs at the level of market segments. Unfortunately, however, there is no reliable and comprehensive classification of retailers into such segments, and so we must construct our own categorization. To do so, we apply Latent Dirchilet Allocation (LDA) (Blei et al., 2003), a popular technique in natural language processing for uncovering hidden group structure in text-based observations. In our case, the latent groups are the market segments, and the observations correspond to the top 10,000 retailers in ODP, where each retailer is represented by the collection of search queries used to find it, excluding navigational queries.

LDA begins by positing that there exist latent topics (market segments) in the data, that each observation (retailer) is an unknown mixture of these latent topics, and that each topic (market segment) corresponds to an unknown distribution over terms (search queries). For each observation, it is further assumed that each term is generated by first sampling a topic from the observation's topic distribution, and then sampling a term from the topic's term distribution. Thus, the model in effect assumes that when a user issues a search query that ultimately results in visiting a retailer, that query is constructed by first probabilistically selecting a market segment (*e.g.*, travel), and then probabilistically selecting a term associated with that segment (*e.g.*, airfare). Though these selection distributions are all *a priori* unknown, LDA efficiently infers them from the data. Ultimately, each retailer is associated with a model-inferred distribution over retail segments. This “mixed membership” representation is especially useful for large retailers, such as Amazon.com, that often compete is multiple market segments.

LDA requires that one specify the number of market segments to infer, which we set to 100. However, as is common in LDA, some topics have effectively the same semantic meaning for our purposes (*e.g.*, topics corresponding to casual and formal clothing), and some topics are effectively meaningless (*e.g.*, a topic that heavily weights “stop words”, such as “the”, “you”, and “it”).[8] To deal with this issue, we manually examined the 100 algorithmically generated topics, and combined and removed topics based on semantic coherence. For example, topics pertaining to televisions and laptops were combined to a single, consumer electronics category.

This process generated 54 market segments. Each retailer is represented by a vector of length 54, with each entry in the vector indicating the percentage of the retailer’s business assigned to the corresponding market. Most retailers have only a few non-zero entries, indicating that they specialize in only a few classes of goods. However, large firms such as Amazon and Ebay, hold market share in many segments, and correspondingly have a number of non-zero entries.. Our inference procedure is based on the assumption that a retailer’s search volume for a given market segment corresponds to its market share. While this assumption is clearly violated in certain instances, on the whole it seems reasonable accurate.

[8] It is also possible that a single topic is in reality mixing two or more distinct topics; we mitigate this possibility by setting the number of topics (100) relatively high.

2.4 Constructing content provider segments

As with retailers, we seek to classify content providers (*i.e.*, non-retailers) into various categories, such as news, games, and education. As before, existing classifications are insufficient for our purposes, and so we turn to LDA, inferring site groupings via the search queries associated with each website. In this case, we started with 200 LDA topics, and then collapsed these into 31 categories. In constructing the content provider segments, however, we encounter three additional complications. First, our dataset includes over 20 million non-retail domains, many of which were visited only a handful of times, and in particular are associated with relatively few search queries. Such sparsity introduces considerable noise into the LDA classification process, and so we restrict our classification analysis to the 30,000 most visited non-retail domains, which in aggregate account for 84% of (non-retail) web traffic. (For the parts of our analysis that do not require content providers to be classified, we use the full set of non-retail domains.) Second, unlike for retailers, some of the largest content providers often have subdomains that fall into substantively different categories. For example, `google.com`, `mail.google.com`, and `news.google.com` correspond to search, mail and news, respectively. Thus, for Google, MSN, Live, Yahoo and AOL, we classify sites at the level of subdomains; for the remaining sites, we classify them according to their top-level domain. Third, many of the most popular sites exhibited poor classification accuracy, as the search queries associated with them were often not good representations of their general category. For example, "gmail login" was one of the most popular search queries issued for Gmail, providing only limited signal. To mitigate this issue, we augmented the algorithmic LDA classification with hand-labeled categories for the 200 most popular sites.

In contrast to our classification of retailers, each content-producing site is assigned to a single category, either the hand-labeled category for the top 200 sites, or the LDA category with the highest weight for the remaining sites. The reasons for this choice are two-fold: first, for the top sites, producing hand-labeled distributions would have been substantially more

Table 1: Classification of content-producing web sites.

Top-level category	**Secondary category**
Web Services	people search, email, games, social, dating, jobs, gambling/games, scam services, travel booking, gambling/lotto, general web services, video streaming, web search
Publishing	news, entertainment/celebrity, gaming, sports, entertainment/tv, life, health, entertainment/music, general publishing, entertainment/other, religion
Reference	weather, general reference, home, community, education, knowledge, government

difficult than simply assigning each site to a single category; and second, content-producing sites are largely narrowly focused, and so mixed classifications make less sense in this setting. Finally, after examining the resulting web site classifications, we found these could be further grouped into one of three major categories: services (*e.g.*, email and search), publishing (*e.g.*, news), and reference (*e.g.*, education and government). The resulting two-level taxonomy is presented in Table 1.

3 Reliance on Third-Party Capable Advertising

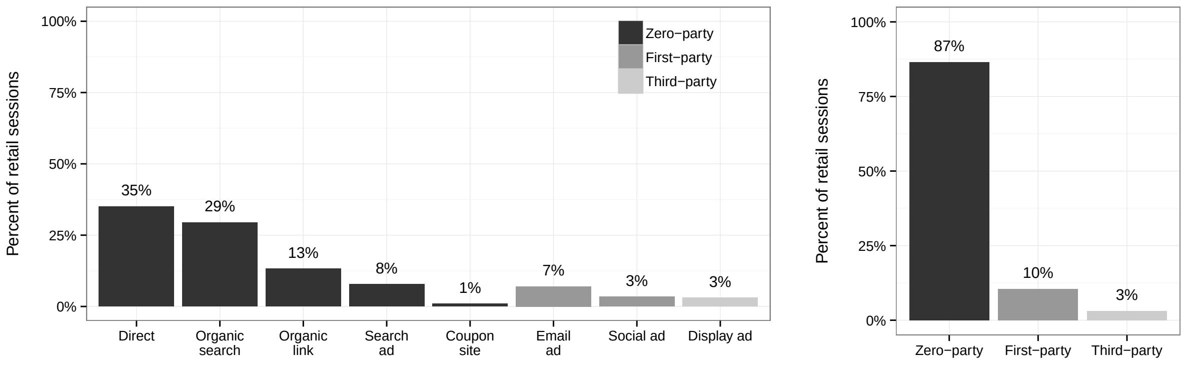
We begin our empirical analysis by examining the potential impact of do-not-track on retailers (Section 3.1) and then consider the effect on content providers (Section 3.2).

3.1 Retailer-centric analysis

As noted in the previous section, for each of the 320,889,786 shopping sessions in our data, we determined the proximate path through which users arrived at the retailer as: direct navigation, organic search, organic website link, search ad, coupon site, email marketing, social ad, or display ad. We now further classify each of these eight possible entry points according to the user data involved: “zero-party,” “first-party,” or “third-party.” Zero-party encompasses instances in which data on a user’s past actions are not directly involved in prompting the shopping session. Direct navigation falls into this category, as does clicking on an organic website link, or a link displayed on a coupon site. Moreover, since both organic search results and search ads are based primarily on the search query, we likewise classify these as zero-party information paths.[9] We label as first-party those instances in which users are targeted for advertising based only on their past interactions with the entity delivering the ad. In particular, social ads (*e.g.*, ads appearing in the Facebook newsfeed) are typically targeted based on actions that users take on the social network itself, such as joining a group or endorsing a product. Similarly, since U.S. law restricts unsolicited email, email marketing typically requires an existing relationship between the customer and retailer, and so is also primarily based on first-party information. Finally, as we have described above, third-party comprises cases where users are targeted based on information that they did not directly provide to the entity displaying the ad. Of the eight paths detailed above, only display ads, which are primarily served via real-time auctions, fall into this category. In fact, many such ads do not use third-party data, instead relying on contextual features of the webpage and the overall demographics of site visitors. However, to be conservative in our analysis, we classify all display ads as “third-party”, which is shorthand for “third party capable,” to reflect the fact that nearly all of these ads could reasonably use third party information.

Figure 1 shows the distribution of entry paths to retail sites, categorized by both the specific mechanism (*e.g.*, direct navigation or email marketing), as well as the information type (*i.e.*, zero-, first-, or third-party). The majority of retail sessions are not initiated by advertising but rather by direct navigation (35%) and organic web search (29%), both of

[9] Though search results are personalized to some extent, and hence draw on past user behavior, the overall effects of such personalization are relatively small (Hannak et al., 2013), and we thus elect to classify it as zero-party. While one could reasonable re-classify organic search as first-party, third-party information is certainly not involved, and so the bulk of our analysis and conclusions remain unchanged.



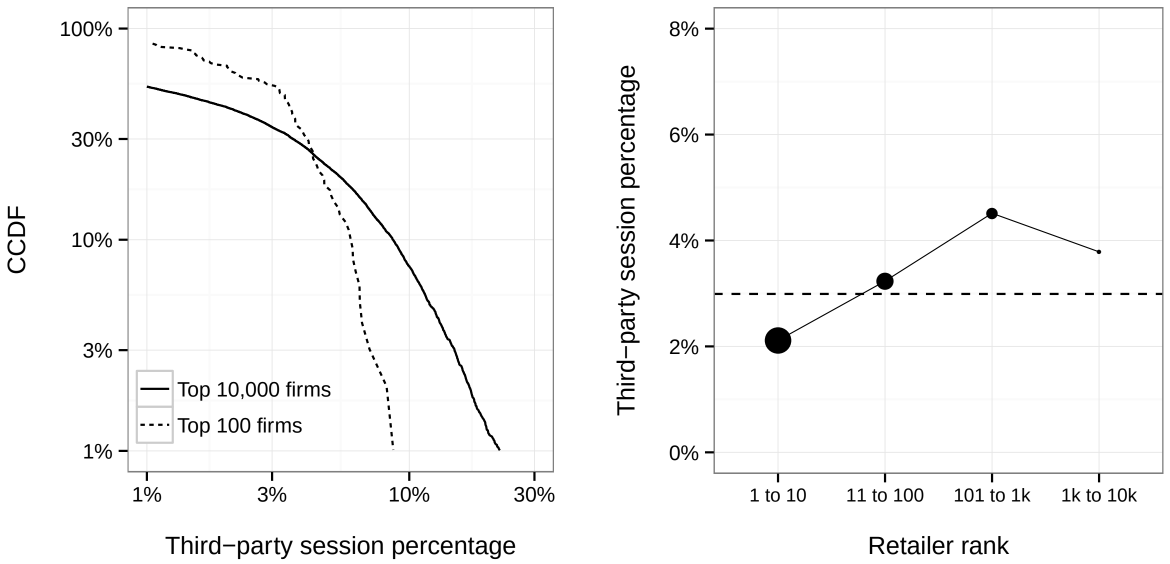
(a) (b)

Figure 1: Starting points for shopping sessions by (a) traffic channel and (b) information type. The majority of traffic to retail sites is driven by direct navigation, organic and sponsored web search, organic web links, and email marketing, none of which use third-party tracking data. Remarkably, third-party advertising accounts for just 3% of shopping sessions.

which are initiated independently by the user. Among advertising channels, email marketing (7%) and sponsored search (8%) dominate, neither of which rely on third-party information. Display advertising initiates only 3% of retail sessions. As summarized in Fig 1(b), nearly all retail sessions are triggered not by third-party data, but by either zero-party (87%) or first-party (10%) information.

Why are only a small proportion of retail sessions initiated by third-party advertising? Though we can only speculate, a likely factor is that the dominate entry mechanisms—direct navigation, organic search, and search advertising, which together trigger 72% of retail sessions—are the result of users actively seeking to visit a retailer. In contrast, display advertising is paired with content supporting other activities, such as reading the news, playing a game, or connecting with friends on social networking sites. As such, as is well-known, display ads have extremely low response rates—on the order of 1 in 1,000. We note, however, that without detailed traffic data of the sort we analyze, it would have been difficult to estimate the overall proportion of retail sessions attributable to such ads.[10]

[10] Given the sheer size of the e-commerce market, display advertising is still a multi-billion dollar industry even though it is a relatively small piece of the pie.



(a)

(b)

Figure 2: Panel (a) shows the distribution of reliance on third-party advertising among retailers. In particular, none of the top 100 retailers (dashed line) and only 7% of the top 10,000 retailers (solid line) have at least 10% of their shopping sessions coming from third-party advertising. After log binning retailers based on their popularity (*i.e.*, number of shopping sessions), panel (b) shows the percent of shopping sessions driven by third party advertising, where points are sized proportional to the overall amount of traffic each bin of retailers receives, and the dotted line indicates the overall percentage of shopping sessions (3%) driven by third-party advertising.

While third-party capable advertising drives a relatively small overall fraction of retail sessions, it could still be the case that some firms are particularly dependent on third-party ads. A niche clothing store, for example, may not yet have the customer base to garner direct visits, and may also not be highly ranked by search engines; accordingly, they might primarily acquire traffic by identifying potential customers via purchased customer profiles and then targeting them with third-party advertising. For each retailer in our dataset we compute the percentage of shopping sessions triggered by third-party capable advertising. The distribution of third-party ad reliance across retailers is plotted in Figure 2(a), for both the top 100 (dashed line) and the top 10,000 firms (solid line). Very few retailers rely heavily on third-party advertising; in particular, none of the top 100 retailers, and only 7% of the top 10,000 retailers have at least 10% of their shopping sessions coming from third-party capable ads.

We next consider what types of firms are most reliant on third-party capable advertising, focusing first on popularity. One might suspect that smaller retailers benefit more from third-party information than larger firms because larger firms have more advanced internal tracking mechanisms and deeper customer databases. We rank and bin retailers by their number of shopping sessions and compute the fraction of sessions each set of retailers gets from third-party capable advertising. Figure 2(b) shows that smaller retailers do indeed rely on third-party capable advertising more than larger ones. Notably, however, even the least popular set of retailers receives less than 5% of their sessions from this channel.

Next, we consider the extent to which particular market segments benefit from third-party capable advertising. Namely, for each of the 54 retail markets, Figure 3 shows the fraction of sessions driven by third-party capable ads, where points are sized proportional to the size of the market. Though there is some variance across markets—and we again see that smaller markets rely more heavily on third-party advertising—no segment gets more than 6% of sessions from such ads.

To do so, we turn to Yelp, a crowd-sourced local business review site that includes entries for many, if not all, merchants with a physical store, and excludes most online-only businesses. We accordingly assume a retailer has a physical store if and only if it appears on Yelp, and in total 4,561 of the 10,000 retailers we consider meet this criterion.[11] We find that on average, retailers with physical stores receive 3.9% of their shopping sessions from third-party capable ads, whereas the number is 2.3% for online-only retailers. The difference is most pronounced

[11] This classification heuristic is not perfect. In particular, a small number of local businesses list `amazon.com`, or `ebay.com` as their homepages on Yelp because they conduct online sales though these channels instead of a privately owned website. To mitigate the impact of such misclassification, we manually classified the top 100 most popular domains that are listed on Yelp.

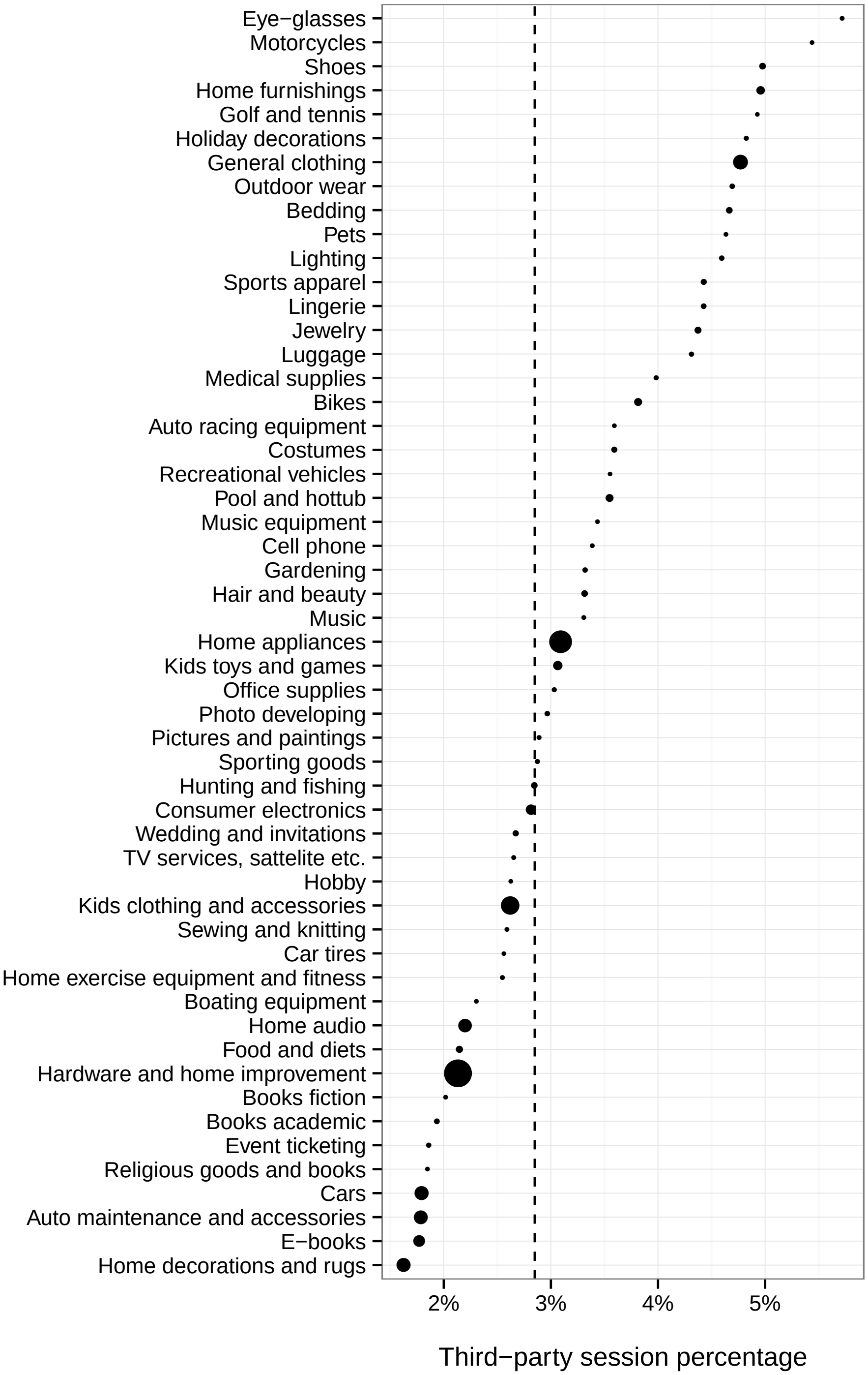


Figure 3: For each of the 54 algorithmically generated categories of retailers, the fraction of shopping sessions driven by third-party advertising, where points are sized proportional the amount of traffic each category receives, and the dashed line indicates the overall average (2.8%).

among the ten most popular retailers in our data. While online-only firms like Amazon and eBay receive a small proportion of their overall traffic from third-party capable ads (1.6%), firms with physical stores like Walmart and Target rely more heavily on third-party capable ads (3.6%). Though this is certainly an important difference from a marketing perspective, in absolute terms the fraction of sessions driven by third-party capable ads is similar small for these two categories of merchants.

Our analysis thus far attributes each shopping session to its proximate entry channel, following a so-called last-click model. While this modeling choice is common in the literature, one might worry that seemingly direct navigation may have in fact resulted from ads a user previously saw. For example, if a user clicks on an ad, but then repeatedly returns to the retailer because of that initial experience, we would misattribute the root cause of those subsequent visits. In this manner we could, at least in theory, underestimate a retailer's reliance on third-party advertising (Lewis et al., 2014). To address this concern, we replace the last-click attribution model with one that takes into consideration the potentially increased consumer awareness, and subsequent visitation, that third-party capable ads can generate. Specifically, for each retail session that we currently classify as direct navigation, we check whether the user visited the retailer via a third-party capable ad within the previous 28 days; if so, we attribute the session to the ad. We find that under this attribution scheme, third-party capable ads account for 3.4% of retail sessions, up from 3.0% using last-click attribution. Thus, while we do see an increase—one that is surely important in the measurement of advertising effectiveness—even this permissive attribution rule suggests that third-party advertising is not a major driver of retail traffic.[12]

A related concern is that conversion rates from clicks on display ads could be higher than that from clicks on other forms of advertising, in which case our analysis would understate

[12] Most ad exchanges that sell ads via "pay-per-conversion" pricing use last-click attribution over a shorter time horizon than a month.

the benefit of display advertising to retailers. Though this is in principle possible, we find that sessions initiated by zero-party channels have, on average, more page views than those from third-party ads, suggesting that display ads do not in fact have higher conversion rates.

3.2 Provider-centric analysis

While we have thus far considered the potential impact of do-not-track on retailers, the perhaps most oft-cited reason against such privacy legislation comes from content providers. Namely, they argue that if web sites could not show highly targeted ads, consumers would have to support web content and services through other, less desirable means, such as micro-payments or subscriptions. Indeed, some of the most visited websites, including Google, YouTube, Facebook and Yahoo, are almost entirely supported through advertising. Their reliance, however, is subject to two caveats. First, as noted above, much online advertising, such as search and social, is not based on third-party data, and would be largely unaffected by do-not-track. Second, there are a number of popular websites—such as Wikipedia and Craigslist—as well as websites for government services, blogs, and personal home pages, that survive, and even thrive, without showing any form of advertising. Thus, the degree to which content providers are truly supported by third-party ads is a subtle empirical question.

Estimating a website's reliance on third-party advertising is difficult since precise revenue breakdowns from advertising and other sources are generally not publicly available. We consequently focus here on simply whether or not a website shows display advertising regardless of how much revenue it earns from those ads. Moreover, though not all display advertising is based on third-party data, we again take the conservative approach and, with two exceptions, call any site that shows display advertising "third-party ad-supported," since the vast majority of display advertising is capable of incorporating third party information through real-time auctions for ad slots.[13] Specifically, we do not categorize display ads on Google

[13] Websites typically show advertising on either over 90% of their pages or on less than 10% of them, and

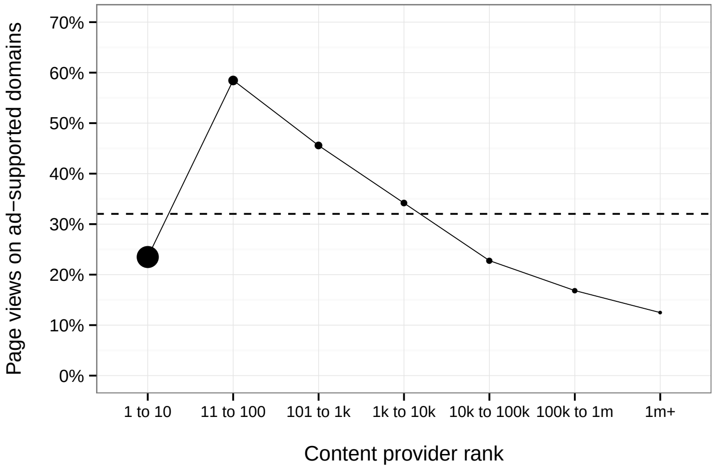


Figure 4: For each set of content providers, log binned by their popularity rank, the percent of traffic from sites that show third-party ads.

or YouTube as third-party since it is known that these ads are targeted based primarily on first-party information, and, given the popularity of these sites, mislabeling them would qualitatively alter some of our results. Overall, our approach is thus a worst-case analysis that effectively upper bounds content providers' reliance on third-party capable advertising.

We find that sites that show third-party capable ads account for 32% of content provider traffic (note that retailers are not included in this or any of the following calculations). While this is certainly not a small fraction, it does indicate, perhaps surprisingly, that web content is not on the whole primarily supported by third-party capable advertising. To investigate further, we show in Figure 4 how ad support varies with site popularity, where sites are log-binned by their traffic rank. Notably, while use of third-party capable advertising is moderate (23%) among the ten most popular content providers, it is quite a bit larger (58%) for those ranked 11–100, and then falls off for lower ranked sites, with only 12% of traffic to content providers outside the top million using display advertising.

To help explain these empirical results, we note that among the top ten content providers,

so we take a conservative stance and call it third-party ad-supported if at least 10% of its page views have display advertising. Recall that in our taxonomy "social ads" are not counted as "display ads", as they are primarily targeted via first-party data.

only two, Yahoo and Microsoft, display third-party advertising. While it should thus be no surprise that the head of the distribution is not primarily supported by third-party advertising, that observation is rarely made in policy discussions. In the tail of the distribution, meanwhile, content providers get too little traffic to make substantial revenue from advertising. For example, even a site that gets 100,000 page views a month—which would make it moderately successful, ranked in the top 20,000 or so—could expect to earn only a few thousand dollars a year. It consequently makes sense that such moderate benefits are outweighed by the implicit costs of showing ads (*e.g.*, on site design and branding). Finally, in the torso of the distribution (ranks 11–10,000), sites both get enough traffic to make substantial revenue from advertising, but do not have as many monetization options as the largest sites, such as the use of first-party data. We note, however, that while such torso sites do display ads at much higher rates than seen in either the head or tail of the distribution, the majority do not show ads.

As with our analysis of advertisers, we look at how use of third-party capable advertising among content providers varies by market segment. For each of the 31 algorithmically generated market segments, Figure 5 shows the fraction of traffic that is supported by third-party capable ads, where points are sized in proportion to the traffic received by the corresponding market segment. The plot illustrates several striking facts. First, web services—such as search and social networking—which account for 54% of non-commerce page views are by and large not supported by third-party ads, with only 20% of their page views being on third-party ad-supported domains. Web search, for example, is supported by zero-party ads; and the largest social networking site, Facebook, relies on first-party ads. However, email and games—also in the services category—do appear to be generally supported by third-party capable advertising, with about 60% of page views in those two categories being on third-party ad-supported domains. Interestingly, the subcategory of services that most often shows

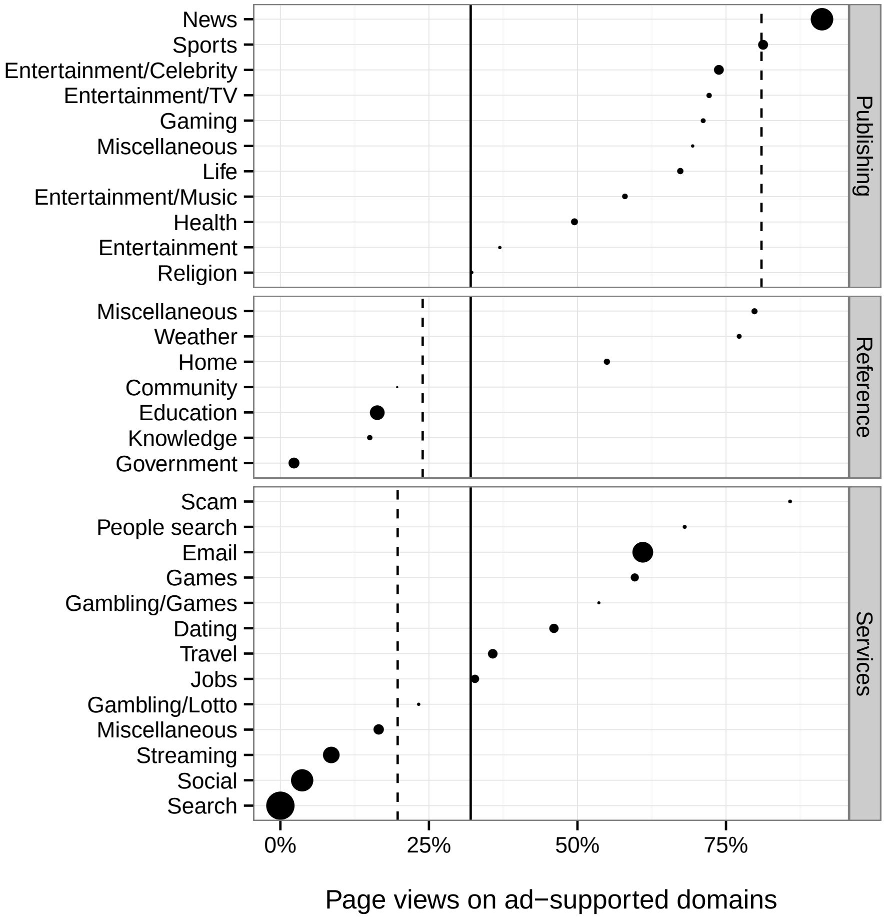


Figure 5: For each algorithmically generated category of content providers, we plot the fraction of page views on domains that show third-party advertising. Points are sized proportional to the amount of traffic each category receives, the solid line indicates the overall average (32%), and the dashed lines indicate the within-group averages.

third-party capable ads (86%) consists of fraudulent sites, such as `mywebsearch.com`.[14] Second, the reference category likewise exhibits only moderate (24%) overall use of third-party capable ads, as many of these sites are not-for-profit, including Wikipedia (in the education category), and various government sites. Among reference sites, weather and general

[14] MyWebSearch is a malicious browser toolbar that users can unwittingly install on their computers if they visit malware-infested websites. Malicious programs like MyWebSearch take control of computers they are installed on, commonly setting themselves as the default search engine and the default homepage on victims' computers, and generate revenue by displaying ads at every opportunity.

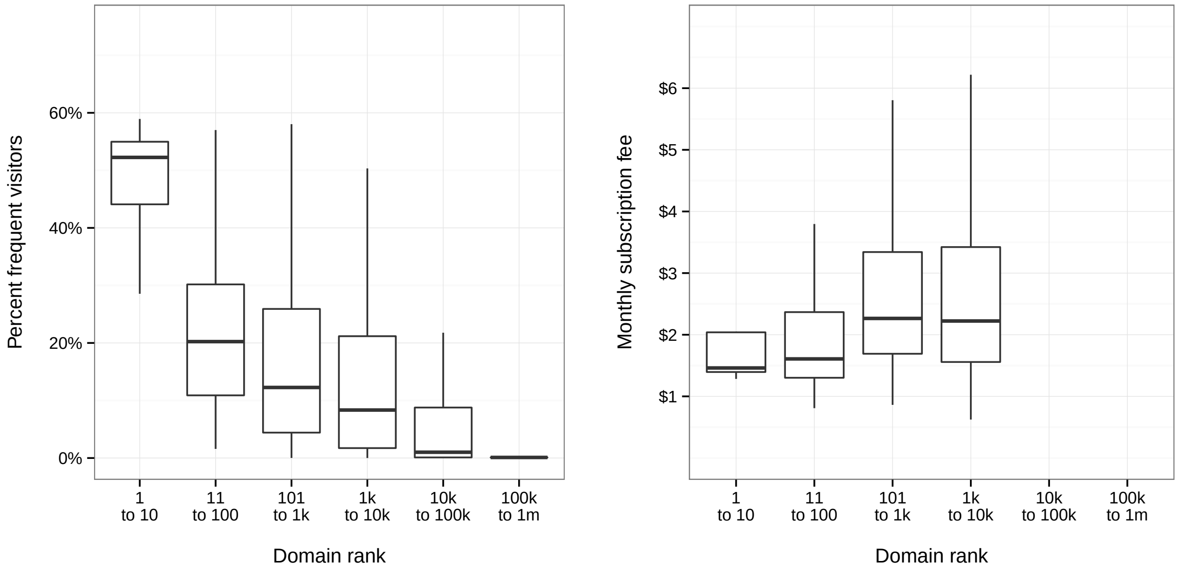
reference (*e.g.*, `ehow.com` and `dictionary.com`), most often show third-party capable ads, with about 75% of traffic in both subcategories accounted for by third-party ad-supported sites. Finally, and most alarmingly, traditional web publishing (*e.g.*, news, sports, and entertainment) is almost entirely display ad-supported (81%). In particular, within the news subcategory—which includes major websites such as Yahoo and MSN—91% of traffic is supported through this channel. Thus, while the majority (68%) of web traffic is not supported by third-party capable ads, certain categories of sites, especially news sites, nearly always is, and could accordingly be substantially impacted by do-not-track legislation.

4 User Browsing Patterns and Paywalls

In contrast to retailers, a substantial fraction of content providers do appear to be at least partially supported by third-party capable advertising. Such advertising, however, is not the only means for generating revenue. For example, as noted above, demographic and contextual advertising does not require third-party information, though such options would likely result in some loss of advertising effectiveness, and thus revenue. Consequently, here we consider the feasibility of one particularly popular alternative to advertising: metered paywalls, or the "freemium" model, in which a site offers its content for free to users who only intermittently visit, but charge a subscription fee to its most loyal consumers, who ostensibly place a relatively high value on its content. Such a payment scheme has in fact already been employed by many major newspapers in the U.S., including the *New York Times* and the *Wall Street Journal.*[15] A related approach is to offer exclusive content for subscribers, which is employed by *ESPN* and many daily newspaper published by the Hearst Corporation.

We begin by estimating the fraction of each ad-supported site's audience that is "loyal,"

[15] Subscriptions account for the majority of revenue of these two newspapers.



(a)

(b)

Figure 6: For content providers, log binned by their popularity rank based on the number of page views they receive, Panel (a) shows the fraction of monthly users that are frequent visitors (*i.e.*, visit a site at least 10 times per month on average). All but the lowest ranked sites have a non-negligible fraction of loyal users. Panel (b) shows the estimated amount that subscribers would have to pay in monthly fees to generate revenue equal to that from third-party capable advertising, where we assume 25% of loyal users subscribe to the site. Note that for bins 10K–100K and 100K–1M, we can only compute these numbers for the small percentage of sites that have non-zero loyal visitation, and so do not show these estimates.

where we define a user as loyal if he or she visits the site at least 10 times per month on average during our 12-month observation period.[16] In Figure 6(a), we bin third-party ad-supported websites by their popularity, and then plot the relationship between a site's popularity and its fraction of users that are loyal. Among the top ten third-party ad-supported websites, a relatively large proportion of users are loyal, 55% on average across the ten sites. This fraction, however, falls off quickly with popularity. For example, for sites ranked 1,000 to 10,000, the median percentage of loyal users is 15%, and sites outside of the top 10,000 almost no measurable loyal visitation. Thus all reasonably popular sites, which are precisely

[16] We restrict our analysis to active users, those who visit at least one web page—on any domain—each month.

the sites that rely on advertising the most, indeed have a non-negligible base of loyal users who could potentially subsidize the site.

While having a base of loyal users is necessary for a freemium monetization model, the key question for assessing feasibility is how much such loyal users would need to pay. This is an admittedly difficult question that depends on a variety of factors, including price elasticity of demand and substitutability for alternative content. Nevertheless, we derive an approximate estimate as follows. First, we assume that each page view generates $0.005 in ad revenue, a bit higher but generally in line with reported estimates (Johnson, 2013). We further assume that 25% of loyal users would be willing to pay a fixed monthly subscription fee, with the remaining loyal users paying nothing, either by limiting their consumption to freely available content or illicitly sharing membership accounts with paying users. While this proportion certainly depends on several important considerations, such as the quality of the content and the subscription fee, an analysis of the New York Times paywall suggests this is a reasonable baseline payment rate (Cook and Attari, 2012). Finally, we compute the monthly subscription fee necessary to offset a site's estimated ad revenue.

Figure 6(b) shows the result of this exercise, plotting estimated subscription fees for sites binned by popularity.[17] For nearly all third-party ad-supported sites ranked in the top 10,000, two dollars a month charged to one-quarter of loyal users is sufficient to offset all ad revenue, largely irrespective of a site's popularity. Sites outside the top 10,000, however, typically do not have enough loyal users to effectively implement a freemium model. Thus though it is hard to definitively assess the feasibility of a freemium monetization strategy, our estimates suggest that for many of the more popular sites, it is an economically reasonable alternative to displaying third-party advertising.

While the subscription fees necessary to offset ad revenue quite can be low, there is still added overhead for registering and paying for sites. Figure 6(a), however, shows that for

[17] We restrict the analysis to sites that have at least one loyal viewer in our dataset.

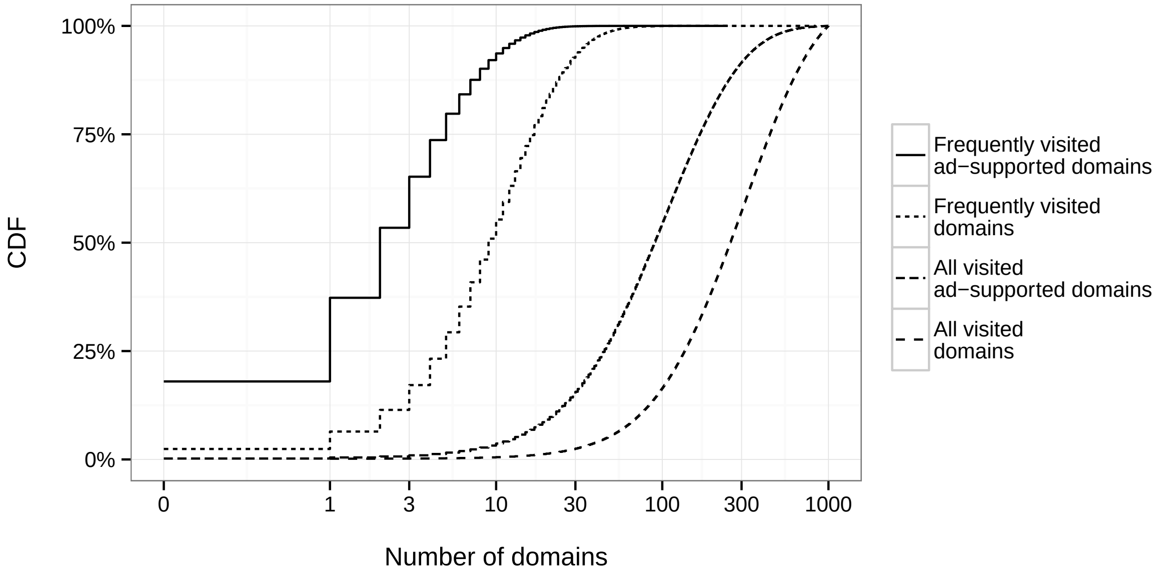


Figure 7: Distribution of the number of domains visited by users, broken down by whether the domain is "susceptible" (*i.e.*, displays third-party capable advertising), and whether the user visited the domain frequently (*i.e.*, at least 10 times per month on average) or simply visited the domain at least once in our six-month observation period. Though most users visited more than 100 domains, they typically only frequently visited 2–3 susceptible domains.

most sites, loyal users comprise just a small fraction of the user base, and thus most visitors would not notice a difference as they would not exceed the free access limits. Still, it could be the case that a non-negligible fraction of individuals regularly visit dozens of third-party ad-supported sites, and so a freemium model could put a heavy burden on these active users. To address this concern, for each user in our sample we computed the number of third-party ad-supported sites they regularly visit (*i.e.*, those they visit at least ten times on average per month). For comparison, we also computed three other statistics for each user: the number of distinct sites they ever visited (regardless of whether the site is supported by third-party capable ads, or whether they visited it regularly); the number of sites they visited regularly (regardless of whether they are third-party ad-supported); and the number of third-party ad-supported sites they ever visited (regardless of whether they visited it regularly). Figure 7 plots the distribution of all four statistics over users. The figure shows that users do visit quite a few sites at least once within the span of a year, approximately 270 on average. Although a handful of major sites tend to dominate overall consumption, people exhibit diverse interests, at least occasionally visiting a number of tail sites, consistent with past work (Goel et al., 2010). If we restrict our attention to third-party ad-supported sites, however, the number falls to about 90. The number of sites users frequently visit is smaller still, 9 on average. Finally, the average number of third-party ad-supported sites users frequently visit is even smaller, just 2; moreover, 95% of users regularly visit no more than 12 such sites. As a result, we expect that under a freemium model users would typically not need to pay for very many sites, limiting the burden such a shift in monetization strategy would place on consumers.

5 Discussion and Conclusion

By analyzing the browsing activity of a large sample of Internet users, we estimated the value of third-party capable advertising for both retailers and content providers. We found that retailers attract only a small percentage (3%) of their customers through third-party capable ads, a result that largely holds across market segments and for firms of vastly different sizes. Looking at content providers, we saw that about one-third of traffic comes from domains that show third-party capable ads, a considerable amount but perhaps smaller than prevailing conventional wisdom. We also found, though, that certain market segments, including news outlets, almost always generate at least some revenue from such ads, making them especially susceptible to privacy policies that limit the use of third-party data for advertising. However, despite the fact that many content providers display third-party capable ads, we showed that the amount they earn from such ads is typically quite modest. In particular, we estimated that small payments by a site's loyal users—conservatively estimated to be on the order of $2 per month—would generate revenue on par with earnings from advertising. Our results thus suggest that privacy legislation like do-not-track would not substantially hamper retailers from attracting customers nor fundamentally reduce the availability of web content and services.

Why is third-party capable advertising not more central to the Internet economy? Though there are a variety of complex dynamics at play, we offer some simple observations. Display advertising by construction is shown out of context—an insurance ad, for example, is shown while reading the news, a banking ad is shown while checking the weather—which explains the extremely low click-through rates of such ads, on the order of 1 in a 1,000. Conversely, when users are interested in shopping, they can directly visit a retailer or search the web, relying on either algorithmic search results or search advertising to guide them. In short, display advertising is unlikely to capture users in the moment at which they are ready to shop and thus even if highly targeted, is not nearly as effective as search advertising.

Further, low click-through-rates for display ads translate to low per page view revenue for content providers. In particular, even moderately successful sites, which might garner 100,000 monthly page views, can expect to earn only a few thousand dollars a year from third-party capable advertising. The vast majority of sites are simply too small for advertising revenue to compete with other motivations for running the site, and we accordingly see limited advertising in the tail. At the other extreme, the largest sites, which could plausibly make considerable revenue from third-party capable advertising, have many more options to monetize their traffic. Google, for example, relies heavily on search advertising, which both yields substantially higher revenue per page than third-party advertising, and does not require any third-party data; Facebook has long avoided paying for third-party data by using its own considerable cache of information on users; the non-profit Wikipedia is so large that it covers its costs by user donations;[18] and Amazon and Ebay sell and facilitate product

[18] Wikipedia's IRS 990 form reveals that the 6th most popular site in the world can be operated for only

purchases. In fact, only two of the top ten sites, Yahoo and Microsoft, earn significant revenue from third-party capable advertising.

What remains is moderately sized websites, which are big enough to earn substantial revenue from third-party capable advertising, but are not big enough to benefit from alternative monetization models or to target users based on first-party information. Of these, high-overhead sites (*e.g.*, news outlets that produce original content), though they often show third-party capable ads, are unlikely to cover much of their costs through such ads. For example, the *New York Times* earns approximately 10% of their revenue from digital ad sales, and many of these ads likely do not use third-party information.[19] Finally, this leaves low-overhead, moderately sized websites (*e.g.*, `weather.com`), which we expect would be directly impacted by do-not-track, and which would ostensibly be forced to operate on lower margins, charge for content, merge with larger sites, or disappear altogether due to competition from free substitutes that do not rely on advertising (*e.g.*, crowdsourced content or government agencies), an assessment in line with recent theoretical work (Campbell et al., 2013).

Throughout our analysis, we have attempted to upper bound the value of third-party data to retailers and content providers. Specifically, we have assumed an extreme case in which privacy policies would eliminate nearly all display ads, even though a substantial fraction of display ads are not based on third-party data and would thus almost certainly be unaffected by legislation such as do-not-track. Moreover, even when ads are targeted based on third-party data, alternative ads could be shown that use only zero- or first-party information. In particular, while our estimates assume that privacy legislation would result in content providers losing nearly all revenue from display ads, the actual revenue loss is likely quite a bit smaller, with recent estimates ranging from 4%–40% (Goldfarb and Tucker,

$25 million a year.

[19] http://www.niemanlab.org/2014/05/the-leaked-new-york-times-innovation-report-is-one-of-the-key-documents-of-this-media-age/

2011a; Johnson, 2013).

Nevertheless, it is difficult if not impossible to fully assess the value of third-party capable advertising to retailers. Brand advertising, for example, is designed to induce later purchases without directly attracting clicks on the ad itself, and so our attribution methodology would miss such effects. We suspect, though, that such potential misattribution does not qualitatively affect our results for five reasons. First, to the extent that channel spillovers from display to other channels like search have been estimated, they appear to be small (Rutz and Bucklin, 2011; Papadimitriou et al., 2011). Second, such misattribution in principle applies to all forms of advertising, including search ads and email ads, dampening errors in the *relative* value of third-party capable advertising in attracting customers, which is our primary quantity of interest. Third, since third-party capable ads directly drive such a small fraction of retail sessions, even quite large misattribution errors are unlikely to qualitatively alter our conclusions. Fourth, brand advertising typically targets a wide range of consumers to raise general awareness, as opposed to being highly personalized, and is thus less likely to rely on third-party information. Finally, as described in Section 3.1, our results are qualitatively similar when we re-categorize direct visits as third-party in cases where the user previously clicked on a third-party ad for the retailer, suggesting that the attribution scheme is not driving the results. It is, however, certainly important to better understand and correct for the effects of misattribution, a task we leave to future work.

A final worry concern is that technological changes could dramatically alter the relative value of third-party capable advertising. For example, with improved targeting tools, retailers may more frequently turn to third-party capable advertising. Similarly, because of attribution errors, it is possible that the market as a whole has undervalued traffic from content providers, who thus stand to earn higher revenue from third-party capable ads at some point in the future. Though such outcomes are certainly possible, given the myriad ways in which online advertising could evolve, we limit our analysis and conclusions to the market in its current form.

In considering privacy policies such as do-not-track, we have analyzed only half the equation, the value of third-party data to advertisers and content providers. In particular, we have not rigorously assessed the benefit to consumers of increased privacy from such legislation. We accordingly cannot offer definitive guidance on whether do-not-track legislation should be enacted or what form it should ultimately take. Nevertheless we close with two reflections. First, content providers have a financial incentive to continue facilitating third-party data collection. Indeed, Facebook, despite their vast amount of first-party information, recently announced their intention to switch from first-party only advertising to allowing the use of third-party tracking data. It thus seems that without legislative action, third-party tracking is likely to increase, for better or for worse. Second, even though the benefits of privacy are hard to quantify, the direct economic gains of tracking are often argued to be so large that they would dwarf any realistic estimate of the value of do-not-track to consumers. Our results, however, suggest that the economic benefits, though ostensibly amounting to billions of dollars, are substantially smaller than generally acknowledged. It is thus possible—though not obvious—that consumer value for increased privacy could tip the scales in favor of enacting protections. Looking forward, we hope our results provide a useful characterization of the drivers of online retail, as well as inform scientific and policy discussions about privacy and the economics of third-party capable advertising.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022.

Broder, A. (2002). A taxonomy of web search. In *ACM SIGIR Forum*, volume 36, pages 3–10. ACM.

Campbell, J. D., Goldfarb, A., and Tucker, C. (2013). Privacy regulation and market structure. *Available at SSRN 1729405.*

Cook, J. E. and Attari, S. Z. (2012). Paying for what was free: Lessons from the New York Times paywall. *Cyberpsychology, Behavior, and Social Networking*, 15(12):682–687.

De los Santos, B., Hortacsu, A., and Wildenbeest, M. R. (2012). Testing models of consumer search using data on web browsing and purchasing behavior. *The American Economic Review*, 102(6):2955–2980.

Deighton, J. and Quelch, J. (2009). Economic value of the advertising-supported internet ecosystem. *IAB Report.*

Farahat, A. and Bailey, M. C. (2012). How effective is targeted advertising? In *Proceedings of the 21st International Conference on the World Wide Web*, pages 111–120. ACM.

Gentzkow, M. and Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839.

Goel, S., Broder, A., Gabrilovich, E., and Pang, B. (2010). Anatomy of the long tail: ordinary people with extraordinary tastes. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 201–210. ACM.

Goldfarb, A. and Tucker, C. (2011a). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404.

Goldfarb, A. and Tucker, C. E. (2011b). Privacy regulation and online advertising. *Management Science*, 57(1):57–71.

Gomer, R., Rodrigues, E. M., Milic-Frayling, N., and m.c. Schraefel (2013). Network analysis of third party tracking: User exposure to tracking cookies through search. In *Proceedings of 2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)*. IEEE/WIC/ACM.

Google (2011). The arrival of real-time bidding. `http://static.googleusercontent.com/media/www.google.com/en/us/doubleclick/pdfs/Google-White-Paper-The-Arrival-of-Real-Time-Bidding-July-2011.pdf`.

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., and Wilson, C. (2013). Measuring personalization of web search. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 527–538.

Johnson, G. (2013). The impact of privacy policy on the auction market for online display advertising. *Available at SSRN 2333193*.

Krishnamurthy, B., Naryshkin, K., and Wills, C. (2011). Privacy leakage vs. protection measures: the growing disconnect. In *Proceedings of the Web*, volume 2, pages 1–10.

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

Lewis, R., Rao, J. M., and Reiley, D. (2014). chapter Measuring the Effects of Advertising: The Digital Frontier. Forthcoming National Bureau of Economic Research Press.

Mayer, J. R. and Mitchell, J. C. (2012). Third-party web tracking: Policy and technology. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 413–427. IEEE.

Papadimitriou, P., Garcia-Molina, H., Krishnamurthy, P., Lewis, R. A., and Reiley, D. H. (2011). Display advertising impact: Search lift and social influence. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1019–1027. ACM.

Reisman, D., Englehardt, S., Eubank, C., Zimmerman, P., and Narayanan, A. (2014). Cookies that give you away: Evaluating the surveillance implications of web tracking. Working paper.

Roesner, F., Kohno, T., and Wetherall, D. (2012). Detecting and defending against third-party tracking on the web. In *In 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI).*

Rutz, O. J. and Bucklin, R. E. (2011). From generic to branded: A model of spillover in paid search advertising. *Journal of Marketing Research*, 48(1):87–102.

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., and Chen, Z. (2009). How much can behavioral targeting help online advertising? In *Proceedings of the 18th International Conference on the World Wide Web*, pages 261–270. ACM.