Do-Not-Track and the Economics of Online Advertising

Ceren Budak  Sharad Goel  Justin M. Rao
Microsoft Research  Stanford University  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 privacy 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 3% of retail sessions are currently initiated by ads capable of incorporating third-party information. Turning to content providers, we find that one-third of traffic is supported by third-party capable advertising, and the rate is particularly high (91%) for online news sites. Finally, we show that for many of the most popular content providers, modest subscription fees of $1-3 per month charged to loyal site users would be sufficient to replace ad revenue. 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, Markus Mobius, David Pennock, David Rothschild, Matthew Salganik and Catherine Tucker for helpful comments and discussion.
1 Introduction

A distinguishing feature of online commerce is the new technologies that link consumers to goods and services. Search engines, for example, let consumers quickly find any available product, search ads connect retailers to individuals expressing specific interests, and algorithmic recommendations help users navigate the plethora of options on retail sites. One such technology that has generated substantial interest in policy circles is “third-party tracking,” which allows retailers to tailor online advertisements using individual-level data collected from around the web. In a common implementation of the technology, ad exchanges and other so-called “third parties” track users as they browse the web in order to build a customer profile (Mayer and Mitchell, 2012); this information is then purchased by, or otherwise made available to, advertisers who bid in real-time display (banner) ad auctions (Google, 2011). In particular, a retailer can “re-target” individuals who have recently visited their site with personalized ads across around the web.\(^1\)

Response rates for such behaviorally targeted ads have been shown to be much higher than ads shown indiscriminately to all users of site. This increased efficiency benefits not only advertisers, but also content providers and advertising exchanges, through higher ad prices (Yan et al., 2009; Goldfarb and Tucker, 2011a; Farahat and Bailey, 2012; Johnson, 2013). Consumers also ostensibly benefit from seeing more relevant ads, and from websites that generate revenue from advertising rather than monetary payments. However, privacy advocates have raised concerns about the practice, proposing “do-not-track” legislation that would limit or prohibit third-party tracking. Despite an active policy debate, there is little rigorous understanding of how third-party tracking fits into the broader Internet economy, and consequently, it is difficult to gauge the potential effects of any such regulation. To shed light on these issues, in this paper we empirically investigate the extent to which third-party

\(^1\)In this case, the retailer has the required data on the consumer but tracking is necessary to identify this user on other websites. This process is known as “cookie synchronization.”
advertising is used by retailers to acquire customers, and by content providers to generate revenue. We further examine—and situate our findings in terms of—other online channels firms use to connect with customers, including web search, sponsored search ads, and ads on online social networks.

In recent years, retailers and content providers have flocked to behavioral tracking, and now over 90% of the top 500 websites share information with hundreds of trackers active in a loosely regulated marketplace for user information (Roesner et al., 2012). Citing the benefits of third-party data for ad effectiveness, advertisers and content providers have largely opposed do-not-track legislation, arguing that such bans would adversely impact their businesses and, in turn, harm consumers, a claim backed by recent academic work. For example, Goldfarb and Tucker (2011b) use survey data to evaluate the impact of the 2002 European Union “Privacy and Electronic Communications Directive” which, among other things, limited third-party tracking. They find that after enactment of the legislation, stated purchasing intent declined on average 65% in the E.U. compared to control countries. Similarly, Johnson (2013) uses auction logs from a real-time display advertising exchange to simulate the impact of privacy policies on ad prices, and finds that a full restriction would reduce prices by about 40%.

Though do-not-track legislation may harm retailers and content producers, it could potentially improve user privacy. Libert (2015) found that 70% of popular health websites leak sensitive information—such as specific conditions, treatments and diseases—to third-party trackers and other firms with which the individual has never directly interacted. This information is typically recorded pseudonymously, for instance using a username or advertising ID, but since the largest trackers have access to browsing behavior across nearly all popular websites (Krishnamurthy and Wills, 2009), the detailed nature of these collected browsing

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%.
histories mean user IDs can usually be linked to names, addresses and other personally identifying information (Krishnamurthy et al., 2011; Reisman et al., 2014). Small trackers, in turn, connect to these large entities to form a small-world network (Gomer et al., 2013), which allows ads to be targeted and delivered with low latency, but also introduces further privacy and security concerns (Reisman et al., 2015).

Past work on the economic consequences of privacy policies (which we have cited above) has largely focused on their impact on ad prices and ad effectiveness. We instead examine the effects of such policies on the Internet economy as a whole, both from the perspective of advertisers and content providers. To do so, we analyze the web browsing histories of 13.6 million users for the 12 months between June 1, 2013 and May 31, 2014. For the advertiser side of the equation, we first identify in our data 321 million visits to the 10,000 most popular e-commerce sites, which we refer to as shopping sessions or retail traffic. We note that our data do not explicitly indicate which shopping sessions resulted in a purchase, though we view such sessions as an important first step in retail transactions. For each shopping session, we then determine the proximate 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. Perhaps surprisingly, display ads—presently the unique form of advertising that could potentially use third-party information—account for only 3% of shopping sessions. Moreover, only 7% of the retailers we study receive at least 10% of their traffic from such third-party capable ads and none of these retailers are in the largest one hundred by overall session volume.

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 such ads. To address this concern, we next examine how reliance on third-party capable ads varies across firms by size, market segment, and offline presence. We find that smaller retailers rely on third-party capable advertising more than larger ones, with the mean moving from
2% in the head of the distribution to 4% in the tail. To examine patterns across market segments, we use topic modeling (Blei et al., 2003) to algorithmically cluster retailers into 54 segments (e.g., sporting goods, home improvement, and books). We find that no market segment receives more than 7% of traffic from third-party capable advertising, with most segments close to the overall mean. Finally, we repeat our analysis separately for online-only retailers, and retailers with a physical store. For the 55% of online-only businesses in our data, we find that third-party capable ads drive 2.3% of their shopping sessions compared to 4.1% for business with both an online and an offline presence. Thus, though there are indeed measurable differences, reliance on third-party capable advertising does not appear to be particularly large across any of these cuts of the data. By way of contrast, we show that this relative uniformity does not hold for reliance on search advertising.

Turning to content providers, we consider the ten million non-retail domains visited by users in our sample. Of these websites, 12% regularly show third-party capable advertising. The sites that do, however, are disproportionately popular, accounting for 32% of aggregate traffic. Outside the top 10,000 sites, relatively few content providers show any form of advertising. 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 use third-party capable advertising. Specifically, 48% of online publishers—accounting for 81% of all online publisher traffic—show third-party capable ads, and among the subcategory of news sites, 91% show such ads.

Given the value of advertising services for the Internet ecosystem (Deighton and Quelch, 2009; Deighton, 2012), and given the benefits of third-party tracking for ad effectiveness, why 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, Reddit
and Wikipedia. Moreover, for smaller sites, the amount they can earn from advertising is relatively modest, suggesting they have other motivations for producing the content.

To better understand the extent to which content providers that show third-party advertising would be adversely affected by privacy policies that limit or ban such ads, we consider their ability to generate revenue via alternative sources. Specifically, we consider one of the most prevalent alternatives to an advertising-only model in the marketplace today: a metered paywall (“freemium”) model, in which site visitors pay subscription fees to consume content in excess of a modest free 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 for successful adoption of this strategy is a set of loyal users who regularly visit the site. Among the top 10,000 sites that show third-party capable advertising, we find that on average 15% of users visit the site at least 10 times per month, with the more popular sites tending to have more loyal visitation. We further estimate that if one-fourth of such loyal users ultimately subscribe to the sites they visit, a monthly fee of $2 would generate revenue comparable to the entire stream from third-party advertising, based on current ad rates (Beales, 2010). And if websites continued to run (first-party) ads, albeit less profitable ones, the required subscription rates would be even lower. Moreover, we find that consumers typically only regularly visit 2–3 sites that feature third-party capable advertising, limiting how much they would need to pay if such sites switched to a freemium model. However, outside the top 10,000 sites, few sites have many loyal visitors, suggesting that a freemium model may not be feasible for these modestly-size content providers. While our estimates are admittedly not based on hard-to-quantify demand elasticities, we do believe they give a general sense of the feasibility of freemium models.

In analyzing the benefits of third-party ads for retailers, we required a way to associate shopping sessions to the channels that fundamentally drove them. This “attribution problem” is widely regarded to be one of the most difficult in the field. In our primary analysis,
we followed standard practice and used the “last-event attribution model” (i.e., we associated each shopping session with the most recent event in a user’s browsing history that preceded it). This modeling choice raises two concerns. First, ads can have an effect in the absence of a click by raising “brand awareness” (e.g., driving direct navigation in the future), or because the impact occurs at a retailer’s brick-and-mortar store (Lewis and Reiley, 2010). Second, ad clicks may not reflect a causal increase in traffic, because the user would have navigated to the retailer anyway by other means (Blake et al., 2014). These two possibilities potentially bias our results, though in opposite directions—in the first case, clicks understate the impact of advertising, while in the second case, clicks overstate impact.

There is unfortunately no error-free attribution scheme. We do, however, address these above concerns to the extent possible with our data. In particular, to examine whether increased brand awareness results in future direct navigation, we consider the following, more conservative click-attribution model: we take all shopping sessions initiated by direct navigation and look four weeks back into a user’s browsing history; if, during that 28-day window, the user clicked on an ad for the retailer, we credit the downstream 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 the effectiveness of their advertising, it is a relatively small difference in the context of overall traffic. Moreover, it is well-known that smaller, niche segments (e.g. religious goods and costume supply, in our categorization) are not typically associated with brand advertising. Our attribution schemes are thus ostensibly more reliable for these segments, yet we do not see substantially higher dependence on third-party capable advertising in these markets. More generally, although we do observe some heterogeneity in reliance on third-party capable advertising—some of which could be because the attribution models are better suited for certain firms—the overall uniformity of reliance, despite the diversity of retailers in our data, is reassuring.
Do-no-track, along with net neutrality (Economides and Hermalin, 2012) and the collection sales tax (Einav et al., 2012), is one of the most important pieces of legislation concerning the Internet economy, and we view our results as a useful data point for gauging the potential impacts of such regulation. Our empirical analysis, however, considers only one-half of the do-not-track cost-benefit equation, assessing the potential economic benefits of third-party tracking. To offer definitive policy recommendations, one still requires estimates on the value of privacy, as well as a better understanding of how restrictions on third-party tracking would affect data security. Further, our analysis is based on current market conditions, and it is difficult to predict market responses to industry realignment and technological advances (e.g., micropayments, or advances in tracking technology). Finally, there is not a single, well-defined do-not-track policy, but rather a variety of proposals that fall under this umbrella name. Nonetheless, despite these limitations, our results suggest that the benefits of third-party information to retailers and content providers, while certainly not absent, are smaller than generally believed. Accordingly, if the commonly accepted prior was that the economic benefits swamped any reasonable privacy valuations, then our results make the optimal policy less clear. Furthermore, by conducting one of the most comprehensive studies of online shopping and web content consumption patterns, our results provide a basis for ongoing marketing and advertising research.

2 Data and Methods

Our primary analysis is based on web browsing records collected via a toolbar 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

---

3These proposals vary on quite consequential dimensions, 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?
that the browser accounted for 25% of U.S. pageviews.\footnote{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.} Though this is admittedly only a subset of online activity, and in particular, does not include mobile traffic, it is still one of the largest datasets of its kind. Upon installing the toolbar, users can consent to sharing their data via an opt-out agreement. 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 (\textit{e.g.,} \url{http://www.amazon.com}), 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. 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
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 ad-
vertising, email marketing, social advertising, display advertising, coupon (or “deal finder”)

\(^5\) For more on HTTP redirects, see http://www.w3.org/Protocols/rfc2616/rfc2616.html.
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. 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. We categorize shopping sessions originating from social networks—Facebook being the dominant example—as driven 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

---

6The 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.

7Retailers 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.
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”). 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 reasonably accurate.

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 en-

---

8It 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.
counter 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 over 20 million 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 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.

13
Table 1: Classification of content-producing (i.e., non-retail) websites.

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

3 Results

We begin our empirical analysis by broadly examining the various means through which retailers attract customers online, with a focus on advertising that utilizes third-party data. In Sections 3.2 and 3.3, we then consider the content-provider (i.e., non-retail) side of the Internet economy.

3.1 A retailer-centric analysis

As noted in the previous section, for each of the 321 million shopping sessions in our data, we categorized 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.\textsuperscript{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 (\textit{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 the content 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 (\textit{e.g.}, direct navigation or email marketing), as well as the information type (\textit{i.e.}, zero-, first-, or third-party). Perhaps surprisingly, the majority of retail sessions are not initiated by advertising but rather by direct navigation (35\%) and organic web search (29\%), both of which are initiated independently by the user. Interestingly, finding products through traditional search engines seems to have replaced dedicated “shop bots” that were popular a decade ago, and which were credited for reducing price dispersion (Smith, 2002). Advertising channels collectively account for 21\% of site visits, a substantial fraction. Among these channels, email marketing (7\%) and sponsored search (8\%) dominate, neither of which

\textsuperscript{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.
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. Notably, third-party advertising accounts for just 3% of shopping 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.

Though advertising channels as a whole drive a considerable fraction of online commerce, third-party data play a relatively small role in initiating shopping sessions. We can only speculate as to why, but 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 products. Search advertising, for example, allows retailers to target users at the precise moment they have expressed a specific retail interest. In contrast, display advertising is paired with content supporting other activities, such as reading the news, which is a well-known factor in their low response rates (on the order of 1 in 1,000).

We note, however, that 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.

While third-party capable advertising drives a relatively small overall fraction (3%) 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 neither have the customer base to garner
direct visits, nor be highly ranked by search engines; accordingly, they might rely more
heavily on identifying and targeting potential customers based on online profiles compiled
by third-party trackers. To investigate this possibility, for each retailer in our dataset we
compute the percentage of its shopping sessions triggered by third-party capable advertising.
The distribution of third-party ad reliance across retailers is plotted in Figure 2(a) (top
panel), for both the top 100 (dashed line) and the top 10,000 firms (solid line). We find
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. In Figure 2(b) (top panel) we directly examine the
relationship between retailer size and dependence on third-party capable advertising. Smaller
retailers do indeed rely on third-party capable advertising more than larger ones, with the
mean moving from 2% in the head to 4% in the tail.

To provide a richer context, we repeat this analysis for our three other categories of
online advertising—search, social, and email. Notably, the distribution of reliance on search
ads is much more dispersed than for display ads. For example, approximately one-third of
firms in the top 100 get more than 10% of their visits from search ads, whereas none of the
top firms reached this level of reliance for display. Moreover, smaller firms rely considerably
more on search advertising than larger ones—average reliance on search advertising moves
from 6% in the head to 12% in the tail and 10% of firms outside the top 100 rely on search
advertising for more than 30% of their visits. One explanation for this relationship is that
smaller retailers are not as prominently featured in organic search results as are their larger
competitors; search ads thus offer them the ability to compete with larger retailers for the
valuable segment of consumers actively searching for products. Figure 2(a) also reveals that
firms in the top 1,000, except the very largest, are most reliant on email advertising. Social
advertising shows a somewhat similar distribution of reliance to display advertising, but in
Figure 2: Panel (a) shows the distribution of reliance on advertising among retailers across the four major advertising channels. Panel (b) shows the percent of shopping sessions driven by advertising as a function of retailer size (log bins of rank), 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 driven by advertising.
this case larger firms show slightly higher, as opposed to lower, reliance.

Returning to third-party capable advertising, we next consider the extent to which market segments vary in their reliance. 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, no segment gets more than 6% of sessions from such ads. The heterogeneity in reliance we do observe is slightly inversely correlated with market segment size, with smaller markets tending to rely a bit more on third-party advertising, just as small firms did.

3.1.1 Channel attribution

As noted above, a difficult methodological issue with our analysis is accurately attributing retail sessions to the channel that fundamentally drove them. On the one hand, measuring clicks may understate the value of an ad. In particular, brand advertising might drive direct navigation in the future, or ads could generate sales that occur through unmonitored channels, such as in brick-and-mortar stores. On the other hand, clicks may overstate the value of an ad, since users may have visited the site even in the absence of advertising. Lewis and Reiley (2010) provide evidence for the first case by running a field experiment on existing customers of a large department store that primarily does business offline. Blake et al. (2014) provide evidence for the latter by using data from a field experiment on eBay. Our results are thus only imperfect measures of the drivers of retail activity. In particular, the last-click attribution model we use may understate reliance of third-party ads among the largest retailers—who are especially likely to employ brand advertising—and may partially explain the particularly low fraction (3%) of third-party sessions observed for such firms. We observe, however, that even for the smallest retailers, which by-and-large do not engage in brand advertising, only 4% of shopping sessions are driven by third-party capable ads, providing some reassurance that the effects of misattribution do not qualitatively change our
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%).
results. We further note that with the transition to real-time ad exchanges, retailers have increasingly shown preferences for pay-per-click contracts, suggesting that clicks are not as problematic a proxy for value as one might suspect at first glance.\textsuperscript{10}

Nevertheless, to guard against the possibility of misattribution, we conduct three explicit robustness checks. First, we loosen the requirements of the last-click model, replacing it with an attribution scheme that takes into consideration the potentially increased subsequent visitation, for instance due to heightened awareness, that can occur following an ad click. 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 observe a 14\% increase—a magnitude that is surely important in the measurement of advertising effectiveness—the thrust of our results are largely unchanged.\textsuperscript{11}

We next consider whether offline effects confound our analysis. For this 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.\textsuperscript{12} We find that on average, retailers with physical stores receive 3.9\% of their shopping sessions from third-party capable ads, whereas the

\textsuperscript{10}One of the reasons that many advertisers prefer pay-per-click transaction is that it reduces uncertainty about what has actually been purchased. For instance, it minimizes the chance that an ad is served in a hard-to-see place, such as below the fold. It also makes measuring ROI easier, and makes display advertising more comparable to search advertising.

\textsuperscript{11}Most ad exchanges that sell ads via “pay-per-conversion” pricing use last-click attribution over a shorter time horizon than a month. That is, any sale within $X$ days following a click is credited to the ad, if it was the only ad clicked.

\textsuperscript{12}This classification heuristic is not perfect. In particular, a small number of local businesses list \texttt{amazon.com}, or \texttt{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.
number is 2.3% for online-only retailers. The difference is most pronounced among the ten most popular retailers in our data. Whereas 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%). Again, though this is certainly an important difference from a marketing perspective, the fraction of sessions driven by third-party capable ads is similarly small in absolute terms for these two categories of merchants.

Finally, we check whether differences in channel-specific conversion rates skew our results. Specifically, if display ads have higher conversion rates than other paths, our analysis would understate the benefits of display advertising to retailers. Though this is in principle possible, we find the opposite. We see that display ads typically lead to shopping sessions with fewer pages per session than those from zero- and first-party channels. Moreover, it has been previously shown that session depth correlates quite strongly with likelihood of purchase (De los Santos et al., 2012), suggesting that display ads have lower than average conversion rates.

### 3.2 A 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 websites 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. Specifically, we do not categorize display ads on Google 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

\[13\] Websites typically show advertising on either over 90% of their pages or on less than 10% of them, and 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.
content providers outside the top million using display advertising.

To help explain these empirical results, we note that among the top ten content providers, 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 third-party capable ads (86%) consists of fraudulent sites, such as mywebsearch.com.\footnote{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.} 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 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 are, and could accordingly be substantially impacted by privacy legislation.

### 3.3 The feasibility of “freemium” models

As described above, a substantial fraction of content-providers are at least partially supported by third-party capable advertising. Under do-not-track legislation, such sites would likely continue showing ads, though targeted based on site content and overall audience demographics, rather than third-party data. This switch would result in some loss of advertising effectiveness—Johnson (2013) estimates a 40% loss in revenue, although the impact
would vary by site, depending on a variety of factors. For example, a site specializing in political commentary, with weak ties to consumer products, might see more loss of revenue than, say, a publisher of technology reviews. In this section, we consider an alternative to ad-supported content. Namely, we assess the high-level feasibility of metered paywalls (i.e., a “freemium” model), in which free content is offered to users who only intermittently visit a site, but a subscription fee is charged to its most loyal consumers, who wish to consume beyond the free allotment. 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.\textsuperscript{15} In a related implementation, providers set aside premium content available only to subscribers, a strategy employed by ESPN and many newspapers published by the Hearst Corporation. We note from the outset that this analysis is inherently speculative, though we believe it is an informative exercise.

A necessary (though not sufficient) condition for a site to adopt a freemium model is a critical mass of loyal users, as they are presumably a superset of those willing to pay a subscription fee. Thus, as a first step, for each content provider we estimate the fraction of its audience that is “loyal,” 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.\textsuperscript{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. The fraction of loyal visitors, however, falls off quickly with site 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 have almost no appreciable loyal users. Thus, nearly all reasonably popular sites indeed have a large base of loyal users who could

\textsuperscript{15}Subscriptions account for the majority of revenue of these two newspapers.

\textsuperscript{16}We restrict our analysis to active users, those who visit at least one web page—on any domain—each month.
Figure 6: 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), by content provider popularity rank. Panel (b)(top) shows the distribution of the number of domains visited by users, broken down by whether the domain shows third-party capable advertising. Panel (b)(bottom) gives the same breakdown but restricts to sites that are frequently visited (i.e., for each user, how many sites do they frequently visit).

Among the set of loyal users, the decision to subscribe depends on numerous factors, including the availability of substitutes, switching costs, and perhaps most importantly, the actual cost of the subscription. Although we cannot rigorously estimate demand elasticities, our data do facilitate a useful back-of-the-envelope calculation of the general magnitudes in question. First, we assume that each page view generates $0.005 in ad revenue, which is higher but generally in line with reported estimates (Beales, 2010). Second, we 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.\(^{17}\) Under these assumptions, we estimate that for most third-party ad-supported sites ranked in the top 10,000, $2 per month, charged to one-quarter of loyal users, is sufficient to offset all ad revenue (\textit{i.e.}, including revenue derived from ads that do not use third-party tracking). While the exact fee required to offset lost ad revenue varies by publisher, the first and third quartiles of the distribution are relatively tight at $1 and $3 respectively. If publishers continued to run (non-third-party) ads, with a hypothetical 50\% reduction in ad revenue—in line with established estimates—the required subscription fee would be $1 per month.\(^ {18}\) It is worth pointing out that while imperfect, these estimates coincide with the range of subscription fees ($1–$3 per month) of “Google Contributor,” a relatively new program that allows consumers to turn ads off on a small number (5–10) of participating sites in exchange for a small monthly payment.\(^ {19}\)

Under the assumptions above, only modest subscription fees are necessary to offset advertising revenue for any one site. Is it the case, however, that such these fees would be concentrated on a small segment of active Internet users, resulting in prohibitively large payments for any one user? To check, we compute the number of third-party ad-supported sites each user regularly visits. For comparison, we also compute 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

\(^{17}\) Of course the validity of this assumption depends critically on the prices and availability of free substitutes. In the context of the prices we estimate ($1-$3) and based on analysis of customer retention in the New York Times paywall which uses prices that are 10 times higher (Cook and Attari, 2012), it seems like a reasonable baseline rate.

\(^{18}\) Large newspapers like the \textit{New York Times} typically charge more than $10 for digital-only subscriptions. This is much higher than the figure we estimate because newspapers earn very little of their revenue from online advertising. Leaked data on the \textit{New York Times} reveal that less than 10\% of revenue comes from online advertising, despite it being one of the most popular online news sites. Here, subscriptions do not displace ad revenue, but rather the paper’s entire business model is predicated on relatively high-priced subscriptions.

\(^{19}\) See \url{https://www.google.com/contributor/welcome/}.
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 6(b) plots the distribution of all four statistics over users. The figure shows that users visit many sites at least once within the span of a year, approximately 270 on average. This estimate is consistent with past work: although a handful of major sites dominate overall consumption, people exhibit diverse interests, at least occasionally visiting a number of tail sites (Goel et al., 2010). However, if we restrict attention to third-party ad-supported sites, the median falls to 91. The number of sites users frequently visit is smaller still, with a median of 9. Finally, the number of these frequently visited sites that are third-party ad-supported is even smaller, with a median of just 2; moreover, 95% of users regularly visit no more than 12 such sites. It thus appears that most users would not by unduly burdened by a large number of subscription fees.

4 Discussion and Conclusion

By analyzing the browsing activity of a large sample of Internet users, we conducted one of the largest empirical studies of the e-commerce marketplace as a whole. Beyond documenting baseline descriptive facts, our findings and approach help contextualize regulatory policies that would curtail the use of one, or more, evolving advertising technologies. In this space, do-not-track has received far-and-away the most attention from legislators, who have proposed bills in the U.S. House and Senate to protect consumer privacy, and industry members, who have warned about the possible negative impacts of such legislation on the Internet economy. To help inform this debate, we studied the importance of third-party capable advertising to both retailers, who buy ads to attract customers, and content providers, who sell ad space to generate revenue.
We found that retailers attract only a small percentage (3%) of their customers through third-party capable ads, a result that is consistent across firm size and market segment. In contrast, we found that this relative uniformity is not observed for search advertising. 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, browsing patterns reveal that ad revenue can generally be replaced by a small fraction of loyal visitors paying a modest subscription fee, on the order of $1–$2 per month.

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 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 per 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 outside relatively popular sites. At the other extreme, the largest
sites have more options to monetize their traffic. Google, for example, does not run ads on Google.com despite its enormous page view count, preferring to rely on search advertising alone, 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; and the non-profit Wikipedia is so large that it covers its costs by user donations (both in money and in time).\textsuperscript{20} In fact, only two of the top ten sites, Yahoo and Microsoft, earn significant revenue from third-party capable advertising.

What remains is moderately popular websites, those big enough to earn substantial revenue from third-party capable advertising, but not big enough to have attractive alternative monetization models, or to target users based on first-party information. Of these, high-overhead sites (\textit{e.g.}, news outlets that produce original content), though they often show third-party capable ads, are presently unlikely to cover much of their costs through such ads. For example, the \textit{New York Times} earns approximately 10\% of their revenue from all digital ad sales.\textsuperscript{21} These sites by-and-large either already rely on subscriptions or depend on users or guest contributors to volunteer their time to produce content. Of course, if margins are tight, any damage to revenue can impact a business, but the relatively low current reliance on ad revenue at least reduces the likelihood of a major disruption. Finally, this leaves low-overhead, moderately sized websites (\textit{e.g.}, \texttt{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 (\textit{e.g.}, crowdsourced content or government agencies), an assessment in line with recent theoretical work (Campbell et al.,

\footnote{\textsuperscript{20}Wikipedia’s IRS 990 form reveals that the 6th most popular site in the world can be operated for only $25 million a year.}

\footnote{\textsuperscript{21}These figures are from leaked earnings data (http://www.niemanlab.org/2014/05/the-leaked-new-york-times-innovation-report-is-one-of-the-key-documents-of-this-media-age/) and are consistent with their declared traffic totals and ad rates.}

32
Throughout our analysis, we have attempted to make generous assumptions about the value of third-party data to retailers and content providers. In particular, we have generally 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, recent estimates range from 4%–40% loss in ad revenue, depending on the version of do-not-track under consideration (Johnson, 2013).

Given the difficulty of measuring the causal impact of advertising on sales, it is hard to fully assess the value of third-party capable ads 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 fundamentally confound our results for several reasons. First, to the extent that channel spillovers (e.g., from display ads to search ads) have been estimated, they appear to be very 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. For example, in a large field experiment, Blake et al. (2014) showed that clicks on search ads were often short-cuts for direct navigation, and thus do not represent a causal increase in site visits. 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

\footnote{Similarly, re-targeted ads are designed to capture a consumer who recently looked at specific products on a given site; these ads are known to have much higher click-rates than typical display ads, but it’s unclear how many of these people would have returned to the site later even in the absence of an ad.}
general awareness, as opposed to being highly personalized, and is thus less likely to rely on third-party information. This point is strengthened by the fact that we do not observe large differences across market segments or firms, even though the attribution model is likely a better fit for smaller markets and firms, which do not typically advertise their brand. Finally, as described in Section 3.1, our results are qualitatively similar when we re-categorize direct visits as driven by third-party ads in cases where the user previously clicked on a third-party ad for the retailer, suggesting that the specific attribution scheme is not driving our results. Together, these factors lend credence to our qualitative conclusions. We certainly believe, however, that further work is necessary to better understand and correct for the effects of misattribution, an important task that we leave to future research.

Another complicating issue is that technological changes could alter both the benefits of third-party capable advertising and their privacy costs. In particular, with improved targeting tools, third-party capable advertising may become more effective while simultaneously degrading user privacy. Since it is exceedingly difficult to anticipate the myriad ways in which online advertising could evolve, we limit our analysis to the market in its current form, highlighting how our empirical findings inform the policy debate despite their inherent limitations.

Finally, it is important to point out that our work has focused on only half the cost-benefit equation—we have not assessed the benefit to consumers of increased privacy from such legislation. Accordingly, we 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 appear to 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 serving only first-party information, recently announced their intention to switch from serving only first-party

\footnote{For example, ad blocking browser plug-ins have surged in popularity in the last two years, with estimates placing usage at 20\% of total users (PageFair, 2014).}
ads to allowing the use of third-party tracking data for some ad formats. 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—that consumer value for increased privacy could tip the scales in favor of enacting regulation.

---

24 Measuring privacy valuations is difficult because most of the costs are psychological, a well-known barrier to quantitative preference elicitation, and important technological aspects of tracking are poorly understood by consumers (TRUSTe and Interactive, 2011).
References


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


