Understanding Emerging Threats to Online Advertising

CEREN BUDAK, University of Michigan
SHARAD GOEL, Stanford University
JUSTIN RAO, Microsoft Research
GEORGIOS ZERVAS, Boston University Questrom School of Business

Two recent disruptions to the online advertising market are the widespread use of ad-blocking software and proposed restrictions on third-party tracking, trends that are driven largely by consumer concerns over privacy. Both primarily impact display advertising (as opposed to search and native social ads), and affect how retailers reach customers and how content producers earn revenue. It is, however, unclear what the consequences of these trends are. We investigate using anonymized web browsing histories of 14 million individuals, focusing on "retail sessions" in which users visit online sites that sell goods and services. We find that only 3% of retail sessions are initiated by display ads, a figure that is robust to permissive attribution rules and consistent across widely varying market segments. We further estimate the full distribution of how retail sessions are initiated, and find that search advertising is three times more important than display advertising to retailers, and search advertising is itself roughly three times less important than organic web search. Moving to content providers, we find that display ads are shown by 12% of websites, accounting for 32% of their page views; this reliance is concentrated in online publishing (*e.g.*, news outlets) where the rate is 91%. While most consumption is either in the long-tail of websites that do not show ads, or sites like Facebook that show native, first-party ads, moderately sized web publishers account for a substantial fraction of consumption, and we argue that they will be most affected by changes in the display advertising market. Finally, we use estimates of ad rates to judge the feasibility of replacing lost ad revenue with a freemium or donation-based model.

CCS Concepts: • **Information systems→ Online advertising; Electronic commerce;**

Additional Key Words and Phrases: ad blocking, competition, do-not-track, e-commerce, privacy

1. INTRODUCTION

A distinguishing feature of online commerce is how quickly new technologies change the link between consumers, content providers, and businesses selling 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. The display advertising market is currently facing two disruptions. The first is the rise of ad-blockers, which had 45 million monthly active American users in 2015, up 48% year-on-year (Adobe, 2016).[1] The second is "do not track" restrictions on third-party tracking. In a common implementation of ad tracking technology, ad exchanges and other so-called "third parties" monitor 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 ad auctions [Google 2011]. Through this mechanism, a retailer can, for instance, "re-target" individuals who have recently visited their site with personalized ads across around the web.[2]

Driving both of these changes is a concern for privacy. Half of those that use ad-blocking software cite privacy as their primary reason (Adobe, 2016). 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 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 [Englehardt et al. 2015].

Ad-blocking and tracking restrictions primarily affect display advertising, as opposed to search and native social advertising which are not typically blocked and generally do not use third-party information. Ad-blocking can be viewed as a consumer-driven market response, while policies like do-not-track are pursued through legislative or legal channels. Ad-blocking directly limits the ability to show users ad impressions. Tracking restrictions impact the market through reduced ad effectiveness, since response rates for behaviorally targeted ads have been shown to be much higher than ads shown indiscriminately to all users of a site. In both cases, it is harder for advertisers to reach consumers and for content providers to earn advertising revenue [Farahat and Bailey 2012; Goldfarb and Tucker 2011a; Johnson 2013; Yan et al. 2009]. Given current trends, it is possible that these forces could soon halve the value of the display advertising market.[3] There is, however, little rigorous understanding of how display advertising fits into the Internet economy, and thus how these changes might affect the broader online ecosystem.

To shed light on this issue, in this paper we empirically investigate the extent to which display 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. 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 display ads. Perhaps surprisingly, display ads account for only 3% of shopping sessions. Moreover, only 7% of the retailers we study receive at least 10% of their traffic from such ads, and none of these retailers are in the largest one hundred by overall session volume.

While display 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 display ads varies across firms by size, market segment, and offline presence. We find that smaller retailers rely on display 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 display 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 display 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 display 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 display 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. Given the prominence of advertising in the Internet ecosystem [Deighton 2012; Deighton and Quelch 2009], why is it that two-thirds of Internet traffic comes from sites that do not show display 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.

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 display advertising. Specifically, 48% of online publishers, accounting for 81% of all online publisher traffic, show display ads, and among the subcategory of news sites, 91% show such ads. The “torso” of web publishers—those large enough to make a business of it but not so dominant as to serve native ads with first-party data—are thus likely to experience a significant drop in ad revenue. Without additional sources of income, we would expect many such sites to go under. Since the marginal cost of serving a web page is very low, a new equilibrium with fewer, larger publishers that achieve the scale necessary to turn a profit is a natural prediction.

Content providers would, however, likely seek out new sources of revenue, and we use within-person browsing logs to lend insight into publishers’ ability to pursue such strategies. 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 related strategy that is growing in popularity is to simply ask users for donations. A necessary condition for either is a set of loyal users who regularly visit the site. Among the top 10,000 sites that show display 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 show that if one-fourth of such loyal users ultimately subscribe to the sites they visit (typically 2–3 per month), a monthly fee of $2 would generate revenue comparable to the entire stream from display advertising, based on current ad rates [Beales 2010]. This presumed level of support may be optimistic, but this simple calculation does suggest that sites with a loyal following could offset a non-negligible fraction of their ad revenue with modest participation in a freemium or donation-based model. Outside the top 10,000 sites, few sites have many loyal visitors, suggesting that publishers in this set that rely on ad revenue would be most adversely affected.

It bears emphasis that our analysis runs into a fundamental “attribution problem” for advertising, which is widely regarded to be one of the most difficult in the field. In our primary analysis, we follow standard practice and use the “last-event attribution model” (*i.e.*, we associate 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. 2015]. 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 display advertising from 2.8% to 3.4%. While this 21% 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. As a second check, we consider smaller, niche segments (*e.g.*, religious goods and costume supply, in our categorization), which are not typically associated with brand advertising. In these markets, where our attribution schemes are ostensibly more reliable, we do not see substantially higher dependence on display advertising. Closely related to the click-attribution problem is the conversion-attribution problem. In particular, if clicks on display ads are more likely to convert to sales than those on search or social ads, we might underestimate the value of display advertising. To check, we repeat our analysis by excluding "bounce backs" (retailer visits with only one page view), and find that the percentage of shopping sessions attributed to display advertising drops from 3% to 2.7%. This result suggests that display advertising is less—not more—likely to lead to conversions than other channels.

Despite the limitations of our analysis, our results provide empirical grounding to help understand the emerging threats to online advertising from ad-blocking technology and do-not-track policies. Retailers, we find, should be relatively resilient, since display ads drive only a small fraction of their revenue and they can readily turn to other advertising channels. The largest and smallest content providers likewise seem robust to such technological and policy changes, as they either avoid advertising all together or rely primarily on search or social ads. However, we expect that content providers in the torso of the popularity distribution, particularly web publishers, would be significantly adversely affected by these disruptions. We note that we cannot offer concrete policy recommendations, as that would at the very least require estimates on the benefits to consumers of blocking ads and limiting third-party tracking.

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 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. 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" [De los Santos et al. 2012; Gentzkow and Shapiro 2011].

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`), 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.

We note that our analysis does not include mobile traffic. While mobile usage is on the rise, desktop web browsing is still the predominant channel for digital advertising, and the most popular way to purchase goods online. The latest report from the Internet Advertising Bureau estimates that mobile accounts for 25% of total digital ad expenditures [Allen 2015]. In addition, our focus on ad-blocking and do-not-track is primarily relevant for the desktop browising, since mobile usage is increasingly app-based (see [Herrman 2016]). Within apps, do-not-track generally does not apply, due to terms of service agreements a user must agree to in order to use the app; and ad-blockers typically can not block in-app ads.[6]

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 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.[7] 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

Table I: Classification of content-producing web sites.

Top-level category	Secondary category
Web Services	people search, email, games, social, dating, jobs, games, scam services, travel booking, gambling, 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, gov't.

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 exclude emails containing customers receipts, shipping updates, and other non-marketing information via email. 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 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.

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 Dirichlet 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. We provide more details on using LDA to identify retail segments in Section A.2 of the Appendix. The LDA 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.

Constructing content provider segments. As with retailers, existing classifications are insufficient for our purposes, and so we seek to classify content providers (*i.e.*, non-retailers) into various categories, such as news, games, and education. In this case, we started with 200 LDA topics, and then collapsed these into 31 categories. In contrast to our classification of retailers, each content-producing site is assigned to a single category, primarily because content-producing sites are largely narrowly focused, and so mixed classifications make less sense in this setting. We describe our classification procedure in more detail in Section A.3 of the Appendix. After examining the resulting web site classifications, we found these could be further grouped into one of three major

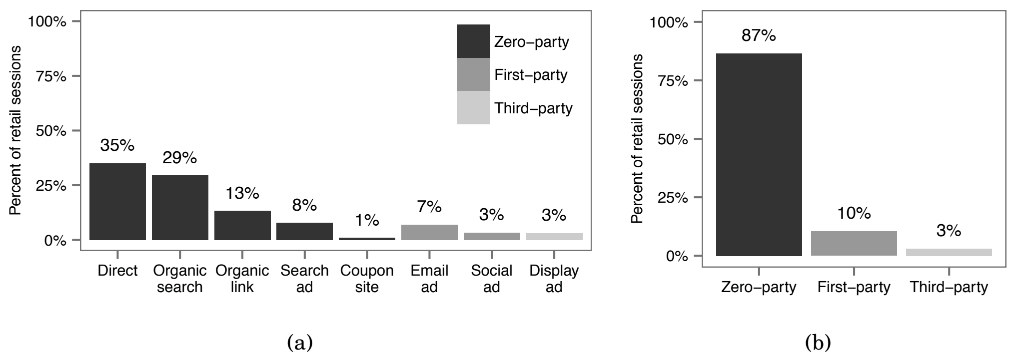

Fig. 1: *Entry points for shopping sessions by (a) traffic channel & (b) information type.*

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 I.

3. RESULTS

We begin our empirical analysis by providing a broad overview of the market from the retailers' perspective, and then dig deeper into the importance of display advertising. We then discuss the content-provider side of the market.

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.[8] 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). 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 rely on third-party information. Display advertising in fact initiates only 3% of retail sessions. As summarized in Fig 1b, 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, display ads 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 display advertising drives a relatively small overall fraction (3%) of retail sessions, it could still be the case that some firms are particularly dependent on these 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 display advertising. The distribution of display ad reliance across retailers is plotted in Figure 2a, for both the top 100 (dashed line) and the top 10,000 firms (solid line). We find very few retailers rely heavily on display 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 display ads. In Figure 2d we directly examine the relationship between retailer size and dependence on display advertising. Smaller retailers do indeed rely on display 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. We also find 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 this case larger firms show slightly higher, as opposed to lower, reliance.

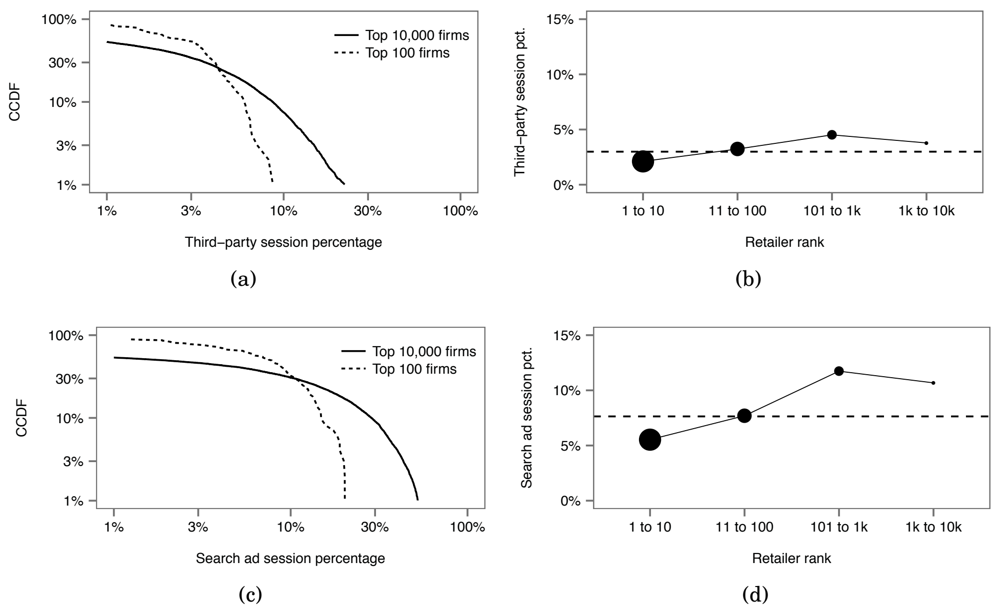

Fig. 2: *Panel (a) shows the distribution of reliance on display advertising among retailers. In particular, none of the top 100 retailers (dashed line) and only 7% of the top 10,000 retailers (solid line) have at least 10% of their shopping sessions coming from display advertising. After log binning 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 display advertising. For comparison, panels (c) and (d) repeat the same analysis for search advertising.*

Returning to display advertising, we next consider the extent to which market segments vary in their reliance. For each of the 54 retail markets, Figure 3 shows the fraction of sessions driven by display 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 display 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. [2015] 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

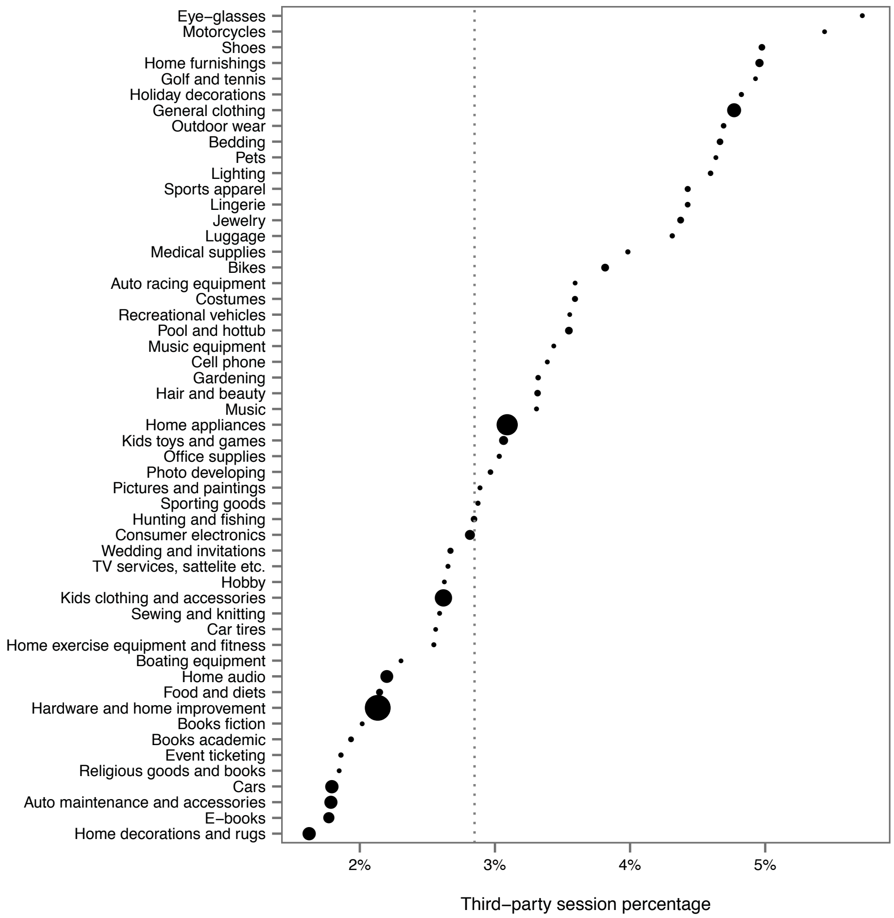

Fig. 3: *For each of the 54 algorithmically generated categories of retailers, the fraction of shopping sessions driven by third-party advertising, where points are sized proportional the amount of traffic each category receives, and the dotted line indicates the overall average (2.8%).*

the largest retailers—who are especially likely to employ brand advertising—and may partially explain the particularly low fraction (3%) of display ad 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 display ads, providing some reassurance that the effects of misattribution do not qualitatively change our 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.[9]

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

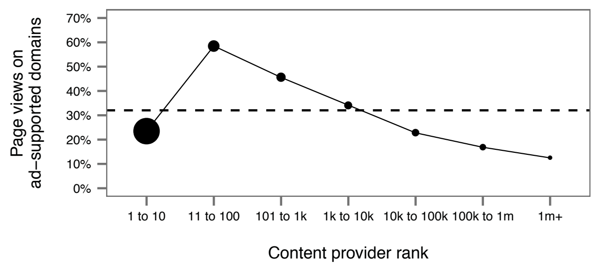

Fig. 4: *For each set of content providers, log binned by their popularity rank, the percent of traffic from sites that are supported by display ads.*

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 display ad within the previous 28 days; if so, we attribute the session to the ad. We find that under this attribution scheme, display ads account for 3.4% of retail sessions, up from 3.0% using last-click attribution. Thus, while we do observe a 13% increase—a magnitude that is surely important in the measurement of advertising effectiveness—the thrust of our results are largely unchanged.[10]

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.[11] We find that on average, retailers with physical stores receive 3.9% of their shopping sessions from display ads, whereas the 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 display ads (1.6%), firms with physical stores like Walmart and Target rely more heavily on display ads (3.6%). Again, though this is certainly an important difference from a marketing perspective, the fraction of sessions driven by display 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. 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. Provider-centric analysis

While we have thus far considered the potential impact of ad-blocking and do-not-track on retailers, the most oft-cited reason against this technology and legislation comes from content providers. They argue that if web sites could not show ads—or were restricted in how they target 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 typically blocked nor is it based on third-party data, and would accordingly be largely unaffected by the disruptions we consider. 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 would be harmed is a subtle empirical question.

Estimating a website's reliance on display 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 and regardless of whether those ads use third-party data. Overall, our approach is a worst-case analysis that effectively upper bounds content providers' reliance on display advertising, and thus bounds their sensitivity to ad blocking software and do-not-track legislation. 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 a site "display ad supported" if at least 10% of its page views have display advertising.[12] We make two exception to this categorization. First, we do not classify YouTube as display ad supported, since the in-video ads shown on the site are not currently blocked by popular ad blocking software—it is technically difficult to do so—and these ads are targeted primarily based on first-party data. Second, we do not classify Google subdomains (e.g., finance.google.com) as display ad supported since Google itself makes the vast majority of its revenue from search advertising. Given the popularity of these sites, mislabeling them would qualitatively alter some of our results.

We find that sites that show display ads account for 32% of content-provider traffic. (We 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 display 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 display 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 showing display advertising.

To help explain these empirical results, we note that among the top ten content providers, only two, Yahoo and Microsoft, are display ad supported. While it should thus be no surprise that the head of the distribution is not primarily supported by display 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 display 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 display ads, where points are sized in proportion to the traffic received by the correspond-

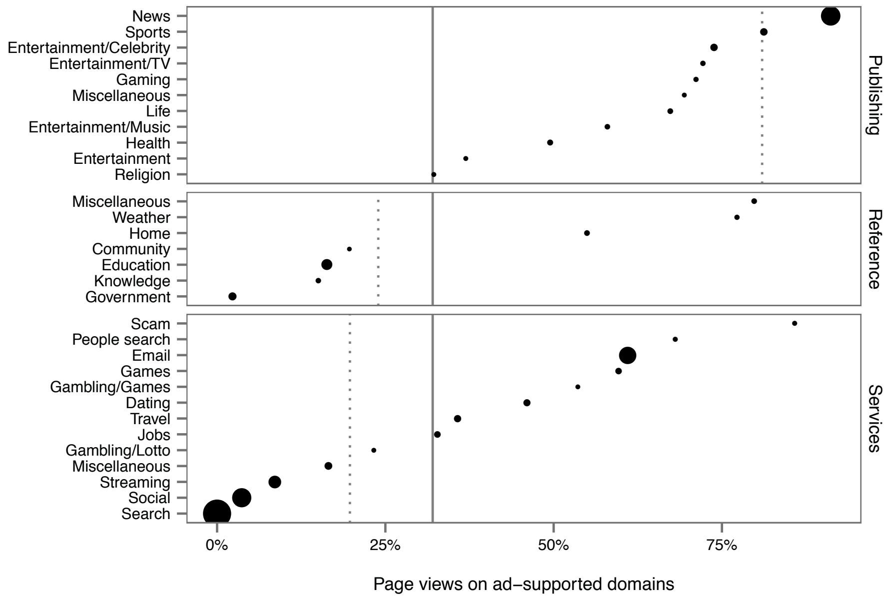

Fig. 5: The fraction of page views on domains that show display advertising, by LDA category. Points are sized proportional to the traffic each category receives. The solid line is the overall average (32%), and the dotted lines are within-group averages.

ing 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 display ads, with only 20% of their page views being on display ad supported domains. Web search, for example, is supported by zero-party ads that are not typically blocked; and the largest social networking site, Facebook, relies on first-party ads that are again not typically blocked. However, email and games—also in the services category—do appear to be generally supported by display advertising, with about 60% of page views in those two categories being on display ad supported domains. Interestingly, the subcategory of services that most often shows display ads (86%) consists of fraudulent sites, such as mywebsearch.com.[13] Second, the reference category likewise exhibits only moderate (24%) overall use of display 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 display ads, with about 75% of traffic in both subcategories accounted for by display 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 display ads, certain categories of sites, especially news sites, nearly always are, and could accordingly be substantially impacted by ad-blocking software and privacy legislation.

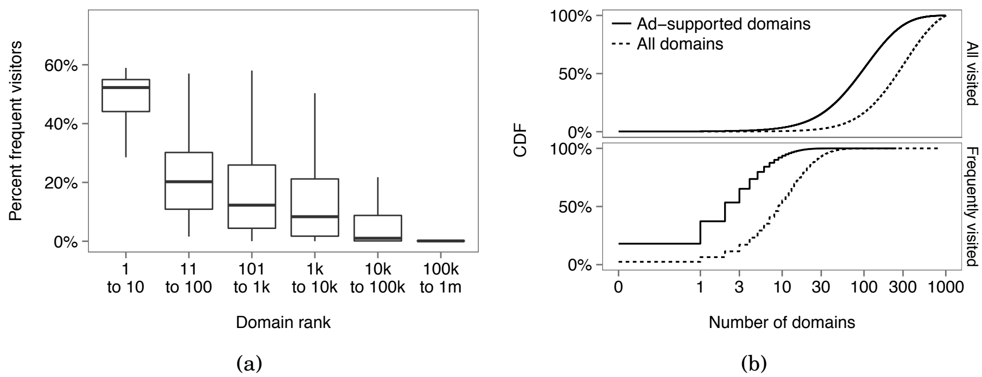

Fig. 6: Panel (a) shows the fraction of monthly users that are frequent visitors (*i.e.*, visit a site at least 10 times per month on average), by content provider popularity rank. Panel (b) (top) shows the distribution of the number of domains visited by users, broken down by whether the domain shows display advertising. Panel (b) (bottom) gives the same breakdown but restricts to sites that are frequently visited (*i.e.*, for each user, how many sites they frequently visit).

3.3. The feasibility of "freemium" models

As described above, a substantial fraction of content-providers are at least partially supported by display advertising. If ad-blocking software were widely adopted, these sites could lose nearly all their revenue from advertising. If do-not-track legislation were passed, 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*.[14] 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.[15] In Figure 6a, we bin display 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 display 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 potentially subsidize the content. Less popular sites lack this base of loyal users, meaning that any subscription scheme, if at all feasible, would likely come in the form of a bundled package, though as we previously showed, such sites are also unlikely to garner much revenue from advertising.

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.[16] Under these assumptions, we estimate that for most display 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. 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.[17] 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.[18]

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 display 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 display ads, or whether they visited it regularly); the number of sites they visited regularly (regardless of whether they are display ad supported); and the number of display ad supported sites they ever visited (regardless of whether they visited it regularly).

Figure 6b 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 display 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 display 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 to date of the online advertising ecosystem. In particular, we have aimed to help retailers, content-producers, and policy makers assess the impact of ad-blocking technology and regulatory policies that limit the use of third-party data for targeted advertising. We find that retailers attract only a small percentage (3%) of their customers through display ads, a result that is consistent across firm size and market segment. Looking at content providers, we see that about one-third of traffic comes from domains that show display ads, a considerable amount but perhaps smaller than prevailing conventional wisdom. We also find, though, that certain market segments, including news outlets, almost always generate at least some revenue from such ads, making them especially susceptible to ad-blockers and do-not-track legislation. However, despite the fact that many content providers show display ads, browsing patterns suggest that ad revenue can generally be replaced by a small fraction of loyal visitors paying a modest subscription fee, on the order of $2 per month.

Throughout our analysis, we have attempted to make generous assumptions about the value of display advertising to retailers and content providers so as to provide a worst-case analysis. In particular, we have generally assumed an extreme case in which privacy policies and ad-blocking technologies would eliminate nearly all display ads. However, given the difficulty of measuring the causal impact of advertising on sales, it is hard to fully assess the value of display 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 small [Papadimitriou et al. 2011; Rutz and Bucklin 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 display advertising in attracting customers, which is our primary quantity of interest. For example, in a large field experiment, Blake et al. [2015] showed that clicks on search ads were often short-cuts for direct navigation, and thus do not represent a causal increase in site visits.[19] Third, since display ads directly drive such a small fraction of retail sessions, even quite large misattribution errors are unlikely to qualitatively alter our conclusions. Finally, as described in Section 3.1, our results are qualitatively similar when we re-categorize direct visits as driven by advertising in cases where the user previously clicked on a display ad for the retailer, suggesting that the specific attribution scheme is not driving our results. Together, these factors lend credence to our qualitative conclusions.

Another complicating issue in understanding the disruptions we discuss is that technological changes could alter both the benefits of display advertising and their privacy costs. In particular, with improved targeting tools, display 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.

Finally, it bears emphasis that our work has focused on only half the cost-benefit equation: we have not assessed the benefits to consumers—-of increased privacy, for example—of ad-blocking software and do-not-track legislation. Accordingly, we cannot offer definitive guidance on whether such technology and legislation should be encouraged 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 serving only first-party 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,[20] the direct economic gains of tracking are often argued to be so large that they would dwarf any realistic estimate of the value of ad-blocking and 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 blocking or regulating display advertising.

Acknowledgments

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

REFERENCES

ALLEN, R. 2015. Mobile internet trends. http://www.smartinsights.com/?attachment_id=53812. Accessed: 2016-05-12.

BEALES, H. 2010. The value of behavioral targeting. *Network Advertising Initiative*.

BLAKE, T., NOSKO, C., AND TADELIS, S. 2015. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica 83,* 1, 155–174.

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. *SIGIR Forum 36,* 2 (Sept.), 3–10.

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. 2012. Economic value of the advertising-supported internet ecosystem. *IAB Report*.

DEIGHTON, J. AND QUELCH, J. 2009. Economic value of the advertising-supported internet ecosystem. *IAB Report*.

ENGLEHARDT, S., REISMAN, D., EUBANK, C., ZIMMERMAN, P., MAYER, J., NARAYANAN, A., AND FELTEN, E. W. 2015. Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. ACM, New York, NY, USA, 289–299.

FARAHAT, A. AND BAILEY, M. C. 2012. How effective is targeted advertising? In *Proceedings of the 21st International Conference on World Wide Web*. WWW '12. ACM, New York, NY, USA, 111–120.

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*. WSDM '10. ACM, New York, NY, USA, 201–210.

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, Washing-

ton, DC, USA.
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*. WWW '13. ACM, New York, NY, USA, 527–538.
HERRMAN, J. 2016. Media websites battle faltering ad revenue and traffic. http://www.nytimes.com/2016/04/18/business/media-websites-battle-falteringad-revenue-and-traffic.html. Accessed: 2016-05-12.
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*. Vol. 2. 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*. WWW '09. ACM, New York, NY, USA, 541–550.
LEWIS, R. AND REILEY, D. 2010. Does retail advertising work: Measuring the effects of advertising on sales via a controlled experiment on yahoo. In *American Economics Association Annual Meeting*. Vol. 3.
LIBERT, T. 2015. Privacy implications of health information seeking on the web. *Commun. ACM 58,* 3, 68–77.
MAYER, J. R. AND MITCHELL, J. C. 2012. Third-party web tracking: Policy and technology. In *IEEE Symposium on Security and Privacy, SP 2012, 21-23 May 2012, San Francisco, California, USA*. IEEE Computer Society, 413–427.
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*. KDD '11. ACM, New York, NY, USA, 1019–1027.
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.
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.
SMITH, M. D. 2002. The impact of shopbots on electronic markets. *Journal of the Academy of Marketing Science 30,* 4, 446–454.
TRUSTE AND INTERACTIVE, H. 2011. Privacy and online behavioral advertising. http://truste.com/ad-privacy. Accessed: 2014-04-10.
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 World Wide Web*. WWW '09. ACM, New York, NY, USA, 261–270.

A. APPENDIX

A.1. Footnotes

Notes

[1]See http://downloads.pagefair.com/reports/2015_report-the_cost_of_ad_blocking.pdf

[2]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."

[3]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. 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%.

[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.

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

[6]One exception is ad-blocking extensions for the Safari iOS mobile web browser, but this accounts for a small fraction of the market.

[7]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.

[8]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.

[9]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.

[10]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.

[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.

[12]In our taxonomy, we recall that "social ads" are not counted as "display ads", as they are not typically blocked and are primarily targeted via first-party data.

[13]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.

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

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

[16]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.

[17]Large newspapers like the *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 *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.

[18]See https://www.google.com/contributor/welcome/.

[19]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.

[20]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].

A.2. Using LDA to construct retail segments

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 the same semantic meaning for our purposes (*e.g.*, topics corresponding to casual and formal clothing), and some topics are meaningless (*e.g.*, a topic that heavily weights "stopwords", such as "the", and "it"). To deal with this issue, we manually examined the 100 algorithmically generated topics, and combined and removed topics based on semantic coherence.

A.3. Using LDA to construct publisher segments

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 LDA classification with hand-labeled categories for the 200 most popular sites. In contrast to our retailer classification, 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; and second, content-producing sites are largely narrowly focused, and so mixed classifications make less sense in this setting.