# Online Advertising Networks

Steven Schmeiser*

Mount Holyoke College

July 11, 2017

**Abstract**

I compile data on the advertising networks used by the top 10,000 websites and use it to present an overview of the online advertising market. I find that 70% of the sites in my sample use a third-party advertising network and that conditional on using at least one, sites use 3.35 on average. Network usage depends on site traffic and category. News and media sites are the most likely to use ad networks and government sites are the least likely. Large general audience sites are more likely to use behavioral tracking technologies than narrow, special interest sites.

*Steven Schmeiser, Economics Department, 50 College Street, South Hadley, MA 01075. Email: steven@schmeiser.org. This research benefited from discussions with Sarah Adelman and Katherine Schmeiser. All errors are my own.

# 1 Introduction

The online advertising market is large and growing. Estimated 2014 revenues were nearly $50 billion in the United States, a 16% increase from 2013 (PricewaterhouseCoopers, 2015). At the same time, the use of third-party advertising networks is coming under scrutiny for privacy, performance, and security reasons. In this study, I collect a new dataset of the top 10,000 websites (as measured by US audience) and the third-party advertising networks that they use. I then analyze the data along several dimensions and document many new findings about the online advertising market.

Rather than directly contract with advertisers, websites often use third-party advertising networks. These third-party networks (hereafter referred to simply as ad networks) contract with and connect advertisers and websites. When a consumer visits a website that belongs to the ad network, the network determines which ad to display to the consumer. In a common scenario, ad selection is done through an auction where advertisers receive information about the impression from the ad network, then place a bid. If the ad network can provide detailed information about the consumer, then advertisers can better target their ads (Yan et al., 2009) and are willing to bid higher prices (Beales, 2010). The ad network can infer information about the consumer based on the website they are visiting (contextual advertising) or by tracking the consumer across all of the websites that belong to the network and building a profile of the user's interests (behavioral advertising). This is an important segment of the economy, as increasing numbers of consumers are turning to digital content for entertainment and news. For example, Pew Research Center (2014) reports that in 2013, 82% of Americans said that they got news on a desktop or laptop computer, and advertising revenue is an important source of income for online news publishers.

However, online advertising networks have come under increased scrutiny for issues around privacy, performance, and security. While regulators have been considering these issues for years (Federal Trade Commission, 2009), recent events such as Apple's decision to allow ad blockers on its mobile web browser have brought discussion about the pros and cons of ad networks into the mainstream. The ability to precisely target ads to consumers relies on the collection and processing of large amounts of consumer data. Evans (2009) and Reisman et al. (2015) discuss the privacy concerns around tracking and behavioral advertising. Slow page load times and large downloads

are often cited as a negative consequence of ad networks.[1] In addition, ad networks may unwittingly spread malware. In 2015, Yahoo's ad network served malicious software to visitors of large web properties.[2]

This paper helps inform the debate about online ad networks by documenting several new empirical facts about the market. Combined, the top 10,000 websites use 561 different ad networks, and on average each site uses 2.35 ad networks. However, many sites do not use an ad network – only 70% use one or more network. Conditional on using advertising networks, sites use 3.35 on average. The number of ad networks is positively correlated with website size as measured by page views, and negatively correlated with page views per unique visitor. I show that different types of websites have different patters of website usage. For example, sites categorized as "News and Media" are much more likely to use ad networks than sites categorized as "Law and Government." In addition, different types of sites use different types of networks. Large general interest sites are more likely to use ad networks that employ behavioral advertising, while special interest sites that cater to a single advertising market are more likely to use contextual networks.

Ad networks are beginning to receive attention in the economics literature. To date, most empirical studies have focused on a narrow or proprietary dataset. Budak et al. (2014) use data collected from the Bing Toolbar browser add-on. The authors use the data to follow a user's behavior and estimate that 3% of retail sessions are a result of ads that incorporate third-party (behavioral) information. The authors also find that between 12% and 58% of content providers show behavioral advertisements, depending on traffic rank. I find that 53.8% of the top 10,000 sites show ads that incorporate behavioral data. While Budak et al. (2014) relies on proprietary data provided by Microsoft, my data is collected using open methods that are available to other researchers. Goldfarb and Tucker (2011a) use data from a media measurement agency to investigate the effects of increasing targeting and obtrusiveness. They find that independently, targeting and obtrusiveness increase the effectiveness of online ads, but that together the effect diminishes. Goldfarb and Tucker (2011b) use data from another field experiment to find that the effectiveness of ads under the jurisdiction of new privacy regulations decreased in effectiveness compared to ads in jurisdictions not under the privacy regulation.

[1] http://www.nytimes.com/2015/08/20/technology/personaltech/ad-blockers-and-the-nuisance-at-the-heart-of-the-mod
html

[2] https://www.washingtonpost.com/news/the-switch/wp/2015/08/04/yahoo-ads-accidentally-spewed-malware/

The most closely related studies are Evans (2009), Gomer et al. (2013), and Roesner et al. (2012). Evans (2009) presents an overview of the online advertising industry. The author provides a summary of how online advertising works, discusses its history, and presents some empirical features. For February 2008, Evans (2009) finds that 56 out of the top 100 websites showed advertisements, and these 56 sites accounted for 77 percent of pageviews (among the 100 sites). In my data, I find that 78 of the top 100 sites use ad networks, and that these 78 sites account for 44 percent of pageviews (again, among the top 100 sites). The lower percentage of pageviews in my data is due to the fact that we measure slightly different things – the display of advertisements versus the use of a third-party ad network. The top two sites in my data (Google and Facebook) do not use ad networks. Instead, these sites use their own advertising platforms. If Google and Facebook are included as using ad networks, the fraction of pageviews using an ad network in my data jumps to 92 percent. Evans (2009) goes on to look at the market structure of the industry. The author lists the top twenty web properties that display ads and ad revenues for some of the properties. In the present paper I present market share statistics for the ad networks themselves in addition to the websites that display the ads.

Gomer et al. (2013) conduct a study that examines the network structure created by third-party networks. The authors collect data on the networks that a consumer is exposed to as they browse the top ten search results (from Google and Bing) to 662 specific search queries. The authors report on the average number of networks a user is exposed to as a function of a website's search rank, the probability that a user is exposed to the top ten third-party networks, and the network structure of the resulting third-party networks. While Gomer et al. (2013) consider all types of third-party networks, including analytics and social media tracking, I restrict my analysis to advertising networks. In addition, I look at the top 10,000 sites, while Gomer et al. (2013) use search results to populate a list of websites to examine.

Roesner et al. (2012) examines the trackers used by the top 500 sites, 500 less popular sites, and simulated browsing sessions based on search terms. They focus on all types of third-party tracking and not just advertising. They find that among the top 500 sites, conditional on having at least one tracker, sites have on average over 7 trackers. In my data, conditional on using ad networks, sites use 3.35 networks on average. The difference is primarily due to my focus on ad networks, a subset of all third-party networks. The authors then go on to break trackers down

into categories based on the technical ways in which they interact with a user's browser. In this study, I examine a larger set of websites and focus only on advertising trackers, and exclude other types such as analytics and social media.

# 2 Data and methodology

## 2.1 Top sites

The hostnames of the top 10,000 sites were collected from the Alexa Top Sites API.[3] The Alexa data includes a website's US rank (as measured by a combination of US unique visitors and page views), global rank, and estimates of reach, page views, and page views per user. Reach is defined as the number of consumers (per million web browsing consumers) that visit the site. Page views per million (PM) are the number of page views (per million total page views) that the site receives, and page views per user (PU) are the number of times that consumers view the site on average. Multiple requests by the same consumer to the same URL on the same day are recorded as one page view.[4] Websites are aggregated to the domain level except for occasional cases where Alexa can distinguish separate sites that use the same domain (for example, different blogs hosted on the same publishing platform). The rank and traffic data reflect a three month period ending March 2016.

## 2.2 Ad networks

Information about ad network usage was collected by visiting each of the 10,000 sites with the Firefox web browser. This was done using the OpenWPM framework (Englehardt et al., 2015). OpenWPM is designed to conduct web privacy studies by using Selenium to automate the process of loading websites with the Firefox browser and routing traffic through a local proxy (MITMproxy) to collect data about the loaded sites. By default, OpenWPM does not record which third party networks a site uses, however, the framework includes the Ghostery browser extension, which can be enabled via a configuration option. The Ghostery plugin displays and optionally

[3] `https://aws.amazon.com/alexa-top-sites/`

[4] Additional details about the Alexa data are available at `https://support.alexa.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined-`.

blocks third party networks loaded by a website. By recording the output of the Ghostery extension, I compiled a list of third party networks used by the top 10,000 sites. I only observe whether or not a site uses an advertising network and not the intensity of the usage. For example, I do not observe how many ads a given ad network displays on a given site. The Ghostery extension also includes basic information about each of the third party networks it detects. Each network has a category (advertising, analytics, tracking, widget, and privacy) and many networks include more detailed information, such as whether or not the network uses behavioral tracking.

The Ghostery identification of behavioral networks is augmented with classifications from the Ad Network Directory and the Network Advertising Initiative (NAI) opt-put registry.[5] The Ad Network Directory maintains a list of advertising networks and includes basic information about each, including whether or not the network uses behavioral targeting. The NAI registry is a list of behavioral networks that let you opt out of receiving targeted ads. I label a network as behavioral if it is marked as such by Ghostery or the Ad Network Directory, or if it is included in the Network Advertising Initiative opt-out list. This method of classifying behavioral networks is imperfect, as it does not directly observe the behind-the-scenes technologies used by a network. In principle, all third-party networks are capable of building a behavioral profile, although not all do. Therefore this method of classification likely underestimates the number of behavioral networks.

My collection methodology records websites' use of all types of third-party networks, such as analytics, social media "share" buttons, and advertising. However, I limit my focus to advertising networks. In the dataset, I label a third-party network as an advertising network if it has a Ghostery category of "advertising" or "tracker," or if it is identified as a behavioral network. I include networks categorized as "tracker" as they are involved with collecting consumer information for advertising purposes even if they do not all display ads themselves.

I loaded each of the 10,000 sites three times over the course of a week. The sites were visited multiple times to catch networks that might not load on every visit and to insure against temporary website downtime. This resulting dataset has 51,343 instances of third-party network–website combinations, 23,500 of which are advertising networks. There are 561 different ad networks used by the top 10,000 websites. 110 of these networks are classified as behavioral networks. To my knowledge, this is the first paper in the economics literature to use automated web crawling

[5] `http://adnetworkdirectory.com`, `http://www.networkadvertising.org/choices/`

to collect information about websites' use of third-party networks.

### 2.3 Website categories

Categorical data is collected from SimilarTech.[6] SimilarTech crawls the web and records the technologies (web server, third party networks, JavaScript frameworks, etc.) that sites use.[7] In addition to recording technologies, they report the category each site belongs to. The SimilarTech data groups websites into 25 top level categories with an additional 190 subcategories under the top level categories. The subcategories often match up to advertising markets such as tennis, golf, cosmetics, weight loss, real estate, and home improvement. SimilarTech's coverage is not complete, and sites can also opt out of SimilarTech's database. Additionally, two sites are categorized as "blocked" and I remove this categorization and include these sites as uncategorized. As a result, only 8,678 of the top 10,000 sites have category data available.

## 3 Descriptive statistics

### 3.1 Websites

Summary statistics for the 10,000 sites are given in the last row of Table 1. Reach and page views PM are not perfectly correlated as consumers visit multiple pages and sites receive different numbers of page view per user. For example, the top site (Google) reaches 84% of consumers but only accounts for 20% of all page views. The distributions of reach and page views PM are highly skewed, with the mean greater than the median. This is due to the power law distribution of web site size, as documented in Schmeiser (2015).

Categorical data is available for 8,678 sites, leaving 1,322 uncategorized. The number of websites and traffic statistics for each of the 25 top level categories (and uncategorized sites) are reported in Table 1. "Arts and Entertainment," "News and Media," and "Shopping" are the largest categories with 1,002, 966, and 903 sites in each category. "Gambling," "Science," and "Pets and Animals" are the smallest, with 36, 41, and 44 sites each.

[6] `https://www.similartech.com`

[7] SimilarTech provides information about the third-party networks that sites use, however I found that many networks were missing from the data. In addition, some companies have multiple ad networks and SimilarTech often aggregates these to the company level rather then the ad network level. For these reasons, I use OpenWPM and Ghostery to collect information about ad network usage.

Table 1: Statistics on the top 10,000 websites. The first column $N$ reports the number of sites in each category. In the following columns, the minimum, median, mean, and maximum statistics are reported for reach, page views per million, and page views per user.

| | | Reach | | | | Page views PM | | | | Page views PU | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | N | Min | Median | Mean | Max | Min | Median | Mean | Max | Min | Median | Mean | Max |
| Adult | 227 | 75 | 232 | 676.029 | 13,000 | 3.040 | 19 | 62.592 | 1,457.200 | 1.010 | 5.850 | 8.426 | 45.800 |
| Arts and Entertainment | 1,002 | 72 | 268 | 1,137.524 | 345,600 | 2.570 | 10.725 | 79.631 | 38,234 | 1 | 2.845 | 4.352 | 65.500 |
| Autos and Vehicles | 171 | 93 | 257 | 475.376 | 8,484.970 | 2.830 | 14.370 | 29.101 | 379.700 | 1 | 4.380 | 5.037 | 30.900 |
| Beauty and Fitness | 85 | 102 | 272 | 603.357 | 5,780 | 2.770 | 15.330 | 42.225 | 495.200 | 1 | 4.600 | 4.879 | 16.520 |
| Books and Literature | 56 | 88 | 277 | 696.536 | 7,810 | 3.510 | 14 | 56.575 | 516 | 1.490 | 4.200 | 5.851 | 41.700 |
| Business and Industry | 765 | 78 | 214 | 538.830 | 22,490 | 2.450 | 10.760 | 35.036 | 4,442 | 1 | 3.980 | 4.983 | 46.600 |
| Career and Education | 804 | 83 | 229.500 | 450.057 | 14,170 | 2.590 | 14.265 | 33.377 | 1,583.800 | 1.040 | 5.400 | 5.844 | 35.600 |
| Computer and Electronics | 613 | 87 | 243 | 733.188 | 31,580 | 2.440 | 10.470 | 32.045 | 1,647.500 | 1.030 | 2.980 | 3.712 | 26.400 |
| Finance | 407 | 97 | 271 | 964.344 | 41,140 | 2.590 | 14.670 | 65.113 | 3,611 | 1.010 | 4.650 | 4.832 | 38.260 |
| Food and Drink | 210 | 96 | 248.500 | 621.667 | 10,847.890 | 2.630 | 10.315 | 27.761 | 369.800 | 1.050 | 3.410 | 3.804 | 17.910 |
| Gambling | 36 | 92 | 198 | 524.661 | 5,510 | 4.310 | 8.810 | 45.627 | 959.900 | 1.010 | 3.445 | 4.755 | 18.900 |
| Games | 302 | 67 | 228.255 | 517.941 | 10,910 | 2.580 | 12.525 | 30.595 | 633.500 | 1 | 3.605 | 5.350 | 60 |
| Health | 260 | 88 | 245.500 | 562.304 | 11,830 | 2.530 | 11.875 | 27.813 | 462.100 | 1.070 | 3.475 | 4.333 | 28.600 |
| Home and Garden | 50 | 95 | 382 | 890.654 | 5,580 | 2.900 | 16.090 | 47.822 | 546.900 | 1.140 | 2.575 | 3.942 | 13.500 |
| Internet and Telecom | 731 | 80 | 293 | 3,340.715 | 839,600 | 2.380 | 12.290 | 445.846 | 198,960 | 1 | 3.110 | 4.139 | 25.400 |
| Law and Government | 252 | 100 | 289 | 611.772 | 12,890 | 2.430 | 13.200 | 34.206 | 963.990 | 1.100 | 3.555 | 3.829 | 10.400 |
| News and Media | 966 | 86 | 355 | 1,606.457 | 203,810 | 2.510 | 10.590 | 64.419 | 19,207 | 1 | 1.920 | 2.723 | 49 |
| People and Society | 220 | 88 | 229 | 574.603 | 11,218.400 | 2.420 | 10.860 | 49.064 | 2,960 | 1.040 | 2.865 | 4.575 | 41.100 |
| Pets and Animals | 44 | 89 | 283 | 431.015 | 1,562 | 2.870 | 13.325 | 24.277 | 130.500 | 1.440 | 4.195 | 4.634 | 15.200 |
| Recreation and Hobbies | 70 | 99 | 211 | 519.611 | 4,640 | 2.900 | 12.230 | 32.062 | 402.200 | 1.250 | 4.515 | 5.143 | 16.160 |
| Reference | 110 | 107 | 417 | 2,592.912 | 122,370 | 2.760 | 13.690 | 113.048 | 5,615 | 1.140 | 2.320 | 3.190 | 13.200 |
| Science | 41 | 116 | 288 | 531.439 | 2,680 | 2.660 | 8.920 | 15.045 | 86.100 | 1.030 | 1.930 | 2.617 | 9.100 |
| Shopping | 903 | 81 | 240 | 1,053.456 | 281,100 | 2.360 | 16.070 | 121.336 | 52,628 | 1.010 | 5.400 | 5.731 | 43.510 |
| Sports | 148 | 85 | 213.500 | 425.032 | 4,900 | 2.800 | 9.545 | 20.375 | 245 | 1.080 | 3.575 | 4.362 | 21.300 |
| Travel | 205 | 95 | 243.620 | 761.899 | 14,940 | 2.810 | 12.570 | 46.541 | 735.600 | 1.010 | 3.860 | 4.298 | 19.800 |
| Uncategorized | 1,322 | 81 | 258.500 | 772.929 | 60,780 | 2.350 | 12.780 | 47.872 | 4,581 | 1 | 3.965 | 4.609 | 58.900 |
| All Sites | 10,000 | 67 | 256 | 1,040.641 | 839,600 | 2.350 | 12.495 | 84.118 | 198,960 | 1 | 3.635 | 4.598 | 65.500 |

### 3.2 Networks

There are 561 different ad networks used by the top 10,000 sites. Only 7,014 of these sites use an ad netework, and I let these sites constitute the market for ad networks. An individual consumer visits many sites and conditional on using ad networks, most sites use more than one. This leads to several different ways of measuring an ad network's reach into the market: the share of page views that use the network, the share of sites that use the network, and the share of consumers that the ad network reaches. My data can address the first two measures, but not the third. For example, if network A is used by sites 1 and 2, each with 50 unique visitors, network A may reach anywhere between 50 and 100 unique consumers depending on the audience overlap of sites 1 and 2.

The first two measures for the top ten ad networks are reported in Tables 2 and 3. These measures differ from a typical market share in that the sum over networks may be greater than one. Each site may include more than one ad network, so in theory two different ad networks could reach every site (or every page view). An alternative measure is the fraction of total network–site combinations, but this is also imperfect, as the theoretical maximum market share for a network is less than one.[8] I use the first measure (fraction of sites and page-views reached) as that is the more relevant measure when thinking about consumers' exposure to networks, and networks' coverage of the market.

Table 2: Ad network reach as measured by page views.

| Rank | Name | Share of page views |
|---|---|---|
| 1 | DoubleClick | 0.512 |
| 2 | ScoreCard Research Beacon | 0.253 |
| 3 | Amazon Associates | 0.180 |
| 4 | Google Adsense | 0.168 |
| 5 | Quantcast | 0.125 |
| 6 | Omniture (Adobe Analytics) | 0.111 |
| 7 | Optimizely | 0.091 |
| 8 | Google AdWords Conversion | 0.083 |
| 9 | Criteo | 0.066 |
| 10 | Adobe Test & Target | 0.061 |

[8]For example, if sites 1 and 2 both use two networks, the total number of network–site combinations is four. However, each network can only reach a maximum of two sites.

Table 3: Ad network reach as measured by sites.

| Rank | Name | Share of sites |
|---|---|---|
| 1 | DoubleClick | 0.397 |
| 2 | ScoreCard Research Beacon | 0.203 |
| 3 | Quantcast | 0.174 |
| 4 | Google AdWords Conversion | 0.149 |
| 5 | Optimizely | 0.135 |
| 6 | Omniture (Adobe Analytics) | 0.114 |
| 7 | Google Adsense | 0.112 |
| 8 | AddThis | 0.080 |
| 9 | Criteo | 0.080 |
| 10 | Amazon Associates | 0.077 |

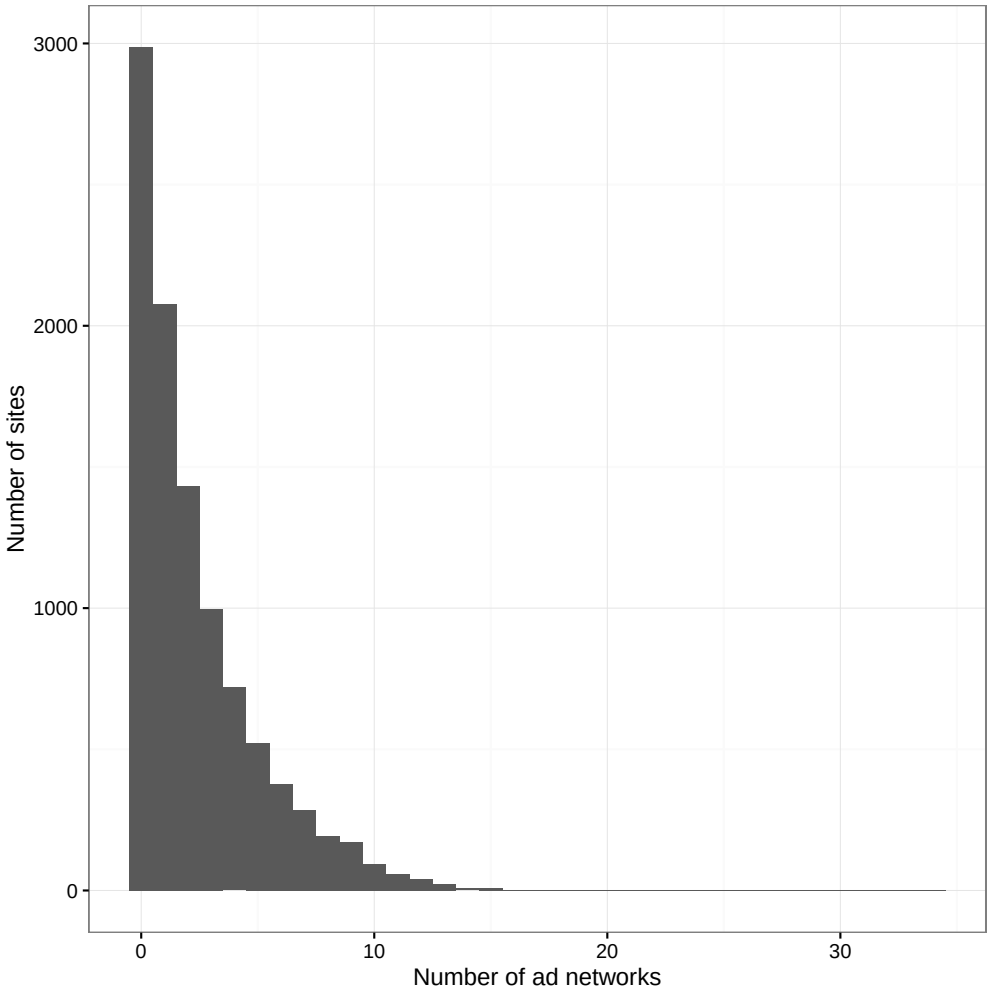
# 4 Network selection

The number of ad networks per site ranges from 0 to 34. On average, the top 10,000 sites use 2.35 ad networks. However, many sites do not display advertisements or use ad networks – for example, product websites, education and government sites, and company homepages. Conditional on using at least one ad network, the mean number of networks per site is 3.35. The distribution of ad networks per site is skewed, with many sites not using any ad networks at all, and a small number using many networks. A histogram of the number of networks per site is shown in Figure 1.

Additional summary statistics, including statistics by site category are shown in Table 4.[9] The first column "Sites" reports the number of sites in a given category. "Share adnet" reports the share of sites in the category that use one or more advertising networks. "Sites adnet," the number of sites in the category that have one or more advertising networks is calculated as "sites" multiplied by "share adnet." The next four columns report the minimum, median, mean, and max number of ad networks per site among sites that have one or more advertising networks. "Share OBA" reports the share of sites (conditional on using one or more advertising networks) that use one or more behavioral networks. The next columns report the minimum, median, mean, and maximum number of behavioral networks per site conditional on the site using advertising networks. I choose to condition the statistics on using one or more advertising networks to eliminate sites

[9] Statistics for the top 100 sites are presented in the Appendix.

Figure 1: Histogram of the number of ad networks per site.



that have no need for an ad network and are not in the online advertising market. In reporting the statistics on the number of behavioral networks, I continue to condition on the use of advertising networks (and not specifically on the use of behavioral networks).

The "News and Media" category has the greatest share of sites using ad networks (93%) and the "Law and Government" category has the least (44%). The "News and Media" category primarily contains sites that publish online content and rely on advertising for revenue. The "Law and Government" category contains many local, state, and federal (.gov) websites that have other means of funding. Conditional on using advertising networks, the "Home and Garden" category uses the most ad networks, with median 5 and mean 5.10 networks per site. The next highest is "News and Media" with median 4 and mean 4.88 networks per site. The "Adult" category uses the fewest networks per site with median 1 and mean 1.96. The "Career and Education" category is the next lowest with median 1 and mean 2.04 networks per site. The "News and Media" category is the most likely to use behavioral advertising (95%) and the "Adult" category is the least likely (52%). This supports the theoretical predictions of Schmeiser (2015) and empirical findings of Budak et al. (2014) that large general audience sites rely more on behavioral targeting than sites with a single advertising market. This is explored further in Section 4.2. The low utilization of behavioral networks in the "Adult" category is also likely due to the increased privacy concerns in this category. The low incidence of behavioral networks in the "Adult" category supports claims that the use of behavioral networks involves privacy trade-offs.

### 4.1 Site size

Here, I explore the relationship between the size of a site and the number of ad networks that it uses. I only include the 7,014 sites that have at least one ad network. Including only the sites that use one network removes from the analysis the decision of whether or not to use ad networks, and isolates the decision of how many ad networks to use. I regress the number of ad networks for site $i$ on the log of reach and the log of page views per user:

$$NUM\_ADNETS_i = \beta_0 + \beta_1 log(REACH_i) + \beta_2 log(PAGE\_VIEWS\_PU_i) + \epsilon_i. \quad (1)$$

The results are shown in Table 5, model 4. The first three models report page views, reach,

Table 4: Statistics on the number of networks per site. The first column "Sites" reports the number of sites in the category. "Share adnet" reports the share of sites in the category that use one or more advertising networks. "Sites adnet" reports the number of sites in the category that have one or more advertising networks. The next four columns report the minimum, median, mean, and max number of ad networks per site among sites that have one or more advertising networks. "Share OBA" reports the share of sites (conditional on using advertising networks) that use one or more behavioral networks. The next columns report the minimum, median, mean, and maximum number of behavioral networks per site conditional on the site using advertising networks.

| | | Share | Sites | Ad networks | | | | Share | Behavioral networks | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Sites | adnet | adnet | Min | Median | Mean | Max | OBA | Min | Median | Mean | Max |
| Adult | 227 | 0.555 | 126 | 1 | 1 | 1.960 | 9 | 0.516 | 0 | 1 | 0.690 | 4 |
| Arts and Entertainment | 1,002 | 0.821 | 823 | 1 | 3 | 3.934 | 34 | 0.818 | 0 | 2 | 1.928 | 16 |
| Autos and Vehicles | 171 | 0.784 | 134 | 1 | 4 | 4.090 | 20 | 0.858 | 0 | 2 | 2.067 | 11 |
| Beauty and Fitness | 85 | 0.835 | 71 | 1 | 3 | 3.535 | 9 | 0.789 | 0 | 1 | 1.648 | 7 |
| Books and Literature | 56 | 0.732 | 41 | 1 | 2 | 2.561 | 13 | 0.707 | 0 | 1 | 1.146 | 6 |
| Business and Industry | 765 | 0.569 | 435 | 1 | 2 | 2.533 | 15 | 0.662 | 0 | 1 | 1.120 | 8 |
| Career and Education | 804 | 0.524 | 421 | 1 | 1 | 2.043 | 10 | 0.703 | 0 | 1 | 1.038 | 7 |
| Computer and Electronics | 613 | 0.626 | 384 | 1 | 2 | 2.747 | 15 | 0.651 | 0 | 1 | 1.188 | 6 |
| Finance | 407 | 0.666 | 271 | 1 | 2 | 2.476 | 11 | 0.697 | 0 | 1 | 1.214 | 6 |
| Food and Drink | 210 | 0.824 | 173 | 1 | 3 | 4.035 | 15 | 0.809 | 0 | 2 | 2.012 | 9 |
| Gambling | 36 | 0.722 | 26 | 1 | 2 | 2.385 | 9 | 0.808 | 0 | 1 | 1.192 | 3 |
| Games | 302 | 0.765 | 231 | 1 | 2 | 3.325 | 15 | 0.814 | 0 | 1 | 1.671 | 8 |
| Health | 260 | 0.700 | 182 | 1 | 2 | 2.846 | 12 | 0.747 | 0 | 1 | 1.401 | 9 |
| Home and Garden | 50 | 0.820 | 41 | 1 | 5 | 5.098 | 13 | 0.976 | 0 | 3 | 2.488 | 6 |
| Internet and Telecom | 731 | 0.513 | 375 | 1 | 2 | 2.821 | 12 | 0.717 | 0 | 1 | 1.363 | 7 |
| Law and Government | 252 | 0.440 | 111 | 1 | 1 | 2.126 | 8 | 0.793 | 0 | 1 | 1.234 | 5 |
| News and Media | 966 | 0.929 | 897 | 1 | 4 | 4.884 | 14 | 0.946 | 0 | 2 | 2.547 | 9 |
| People and Society | 220 | 0.805 | 177 | 1 | 2 | 3.446 | 18 | 0.814 | 0 | 1 | 1.655 | 7 |
| Pets and Animals | 44 | 0.864 | 38 | 1 | 3 | 3.632 | 10 | 0.816 | 0 | 2 | 1.842 | 7 |
| Recreation and Hobbies | 70 | 0.771 | 54 | 1 | 3 | 3.796 | 10 | 0.778 | 0 | 2 | 1.796 | 5 |
| Reference | 110 | 0.745 | 82 | 1 | 2 | 2.841 | 11 | 0.841 | 0 | 1 | 1.524 | 6 |
| Science | 41 | 0.683 | 28 | 1 | 2 | 2.679 | 8 | 0.929 | 0 | 2 | 1.750 | 4 |
| Shopping | 903 | 0.825 | 745 | 1 | 3 | 3.356 | 13 | 0.672 | 0 | 1 | 1.322 | 8 |
| Sports | 148 | 0.899 | 133 | 1 | 3 | 3.602 | 15 | 0.932 | 0 | 2 | 1.932 | 7 |
| Travel | 205 | 0.780 | 160 | 1 | 3 | 3.231 | 17 | 0.819 | 0 | 1 | 1.669 | 7 |
| Uncategorized | 1,322 | 0.647 | 855 | 1 | 2 | 3.200 | 13 | 0.737 | 0 | 1 | 1.504 | 8 |
| All Sites | 10,000 | 0.701 | 7,014 | 1 | 2 | 3.350 | 34 | 0.768 | 0 | 1 | 1.612 | 16 |

and page views per user individually. In model 4, a one percent increase in reach is associated with a 0.318 increase in the number of ad networks used and a one percent increase in page views per user is associated with 0.604 decrease in the number of ad networks. Both coefficients are statistically significant at the one percent level. The total number of page views for a site is the product of reach and page views per user. Model 1 reports that the number of ad networks increases as total page views increase. There results suggest that as a site grows its audience as measured by unique visitors it increases the number of ad networks it uses. However, if page views are driven by the same consumers making repeated impressions, then the number of networks decreases. The incentives driving these patters are an interesting area for future research. In particular, an advertising network can be seen as trying to expand on two margins, the number of ads it serves and the number of unique consumers that are exposed to the network. Serving more ads drives a higher quantity of revenue generating activity, while serving more unique consumers increases the amount of behavioral data that can then be used to better target ads and increase the average price of an ad impression. Increasing page views per user while keeping the number of sites constant only improves the first margin, while expanding the number of consumers reached increases both.

### 4.2 General versus special interest

Here I consider two particular types of sites, general audience and special interest sites. I restrict the set of sites under consideration to those that use one or more advertising networks.

General audience sites cater to a broad set of consumers and a visit to a general audience site does not reveal information about a consumer's interest in a particular advertising market. I classify the top 100 sites in the "News & Media" top level category as general interest sites. Examples include *yahoo.com*, *cnn.com*, and *nytimes.com* at the top end and *techcrunch.com* and *esquire.com* at the bottom ranks. Moving below the top 100 news sites includes many mid-size market local news sites, and I exclude these as the "local" context violates the general audience requirement.

Special interest sites cater to a single advertising market. Therefore a consumer's visit to a special interest site provides specific and useful information to advertisers. Special interest sites provide the information that makes behavioral advertising possible. I designate 34 subcategories

Table 5: OLS regression for number of ad networks per site.

| | Number of ad networks | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| *log(PAGE_VIEWS_PM)* | 0.070*** | | | |
| | (0.027) | | | |
| *log(REACH)* | | 0.358*** | | 0.318*** |
| | | (0.030) | | (0.030) |
| *log(PAGE_VIEWS_PU)* | | | −0.654*** | −0.604*** |
| | | | (0.047) | (0.046) |
| Constant | 3.156*** | 1.245*** | 4.170*** | 2.232*** |
| | (0.081) | (0.179) | (0.066) | (0.192) |
| *N* | 7,014 | 7,014 | 7,014 | 7,014 |
| $R^2$ | 0.001 | 0.020 | 0.027 | 0.043 |
| Adjusted $R^2$ | 0.001 | 0.020 | 0.027 | 0.043 |
| *Notes:* | | ***Significant at the 1 percent level. | | |
| | | **Significant at the 5 percent level. | | |
| | | *Significant at the 10 percent level. | | |

as special interest (listed in Appendix A). Each of the subcategories has between three and twenty sites in the top 10,000. A total of 225 sites are classified as special interest.

Schmeiser (2016) constructs a theoretical model of the choice between behavioral and contextual ad networks and predicts that general interest websites are more likely to use behavioral networks than special interest sites. Special interest sites have the information they need to target advertisements without using a behavioral network, and general audience sites do not. To test this hypothesis, I use the above classification to look for differences in ad network selection between the two types of sites.

First, I consider whether or not sites use one or more behavioral networks. In simple means, 87% of special interest sites use at least one behavioral network, while 99% of general interest sites use behavioral networks. The difference is statistically significant, with a $p$-value of 0.00000114. A Welch two sample t-test has a 95% confidence interval in the difference of the means of (-0.172 -0.074).

Next, I include all sites that use an advertising network and examine the effect of being a

general audience or special interest site on the probablity of using a behavioral network, while controlling for size. I estimate the following model:

$$\begin{aligned} \mathbf{1}\{NUM\_BEHAVIORAL > 0\}_i &= \beta_0 + \beta_1 SPECIAL_i + \beta_2 NOCAT_i + \\ &\quad \beta_3 log(REACH_i) + \beta_4 log(PAGE\_VIEWS\_PU_i) + \epsilon_i. \end{aligned} \tag{2}$$

The *NOCAT* variable is an indicator that the site is not special interest and also not general audience. The $\beta_1$ coefficient on *SPECIAL* is then interpreted as the effect of a special interest site as compared to a general interest site. The results for a logistic regression are reported as model 2 in Table 6. Model 1 in Table 6 reports the results of a logistic regression without controlling for size. I use a logistic regression as the high concentration genreal audience sites that use behavioral networks leads to predicted values greater than one for many general audience sites in a linear probability model. The first model reports the difference without controlling for size. The coefficient on *SPECIAL* is negative at the one percent level, meaning that special interest sites are less likely to use behavioral networks. Controlling for size leaves the point estimate close to model 1, but only significant at the five percent level. The marginal effect of changing from a general audience site to a special interest site (for an average size site, and setting nocat to zero) is negative 0.112 (see Table 7) and is significant at a very low level ($p$-value 0.00002). This closely matches the difference in means presented above of a twelve percentage point reduction in the incidence of behavioral networks between general and special sites.

# 5 Conculsion

I collect a new dataset of the top 10,000 and their choices of advertising networks. I examine the dataset and report several new empirical patterns in the online advertising market, including how the choice of advertising technologies depends on website size and type. The online advertising market is large and growing and has become an important source of revenue for online entertainment and news delivery. The market has also drawn attention from regulators and privacy advocates for its use of large quantities of consumer data. This paper contributes to our understanding of the market as we evaluate the costs and benefits of online advertising, and behavioral

Table 6: Logistic regression for the probability of using one or more behavioral networks.

| | One or more behavioral networks | |
|---|---|---|
| | (1) | (2) |
| *SPECIAL* | $-2.723^{***}$ | $-2.282^{**}$ |
| | (1.021) | (1.026) |
| *NOCAT* | $-3.432^{***}$ | $-3.014^{***}$ |
| | (1.003) | (1.007) |
| *log(REACH)* | | $0.066^{**}$ |
| | | (0.029) |
| *log(PAGE_VIEWS_PU)* | | $-0.491^{***}$ |
| | | (0.043) |
| Constant | $4.595^{***}$ | $4.440^{***}$ |
| | (1.002) | (1.036) |
| *N* | 7,014 | 7,014 |
| Log Likelihood | $-3,765.816$ | $-3,693.334$ |
| Akaike Inf. Crit. | 7,537.631 | 7,396.668 |
| *Notes:* | $^{***}$Significant at the 1 percent level. | |
| | $^{**}$Significant at the 5 percent level. | |
| | $^{*}$Significant at the 10 percent level. | |

Table 7: Marginal effects of the logistic model. The marginal effects are calculated at the mean *REACH* and *PAGE_VIEWS_PU*, but with *NOCAT*=0.

| | d/dx | Std. Err. | z | P>\|z\| |
|---|---|---|---|---|
| *SPECIAL* | -0.112 | 0.026 | -4.270 | 0.00002 |
| *NOCAT* | -0.229 | 0.022 | -10.536 | 0 |
| *log(REACH)* | 0.011 | 0.008 | 1.366 | 0.172 |
| *log(PAGE_VIEWS_PU)* | -0.084 | 0.047 | -1.810 | 0.070 |

advertising in particular.

In addition, this paper contributes to data collection methodologies in economics by using the OpenWPM framework to automate the downloading of data on the technologies used by websites. Research about the online economy presents many opportunities given how much data is publicly available and available for analysis with automated tools.

## References


Beales, H. (2010). The value of behavioral targeting. Technical report, Network Advertising Initiative.

Budak, C., Goel, S., Rao, J. M., and Zervas, G. (2014). Do-not-track and the economics of third-party advertising. *Boston U. School of Management Research Paper No. 2505643*.

Englehardt, S., Eubank, C., Zimmerman, P., Reisman, D., and Narayanan, A. (2015). OpenWPM: An Automated Platform for Web Privacy Measurement. Manuscript.

Evans, D. S. (2009). The online advertising industry: Economics, evolution, and privacy. *Journal of Economic Perspectives*, 23(3):37–60.

Federal Trade Commission (2009). Self-regulatory principles for online behavioral advertising: Tracking, targeting, and technology. Technical report.

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 Schraefel, M. C. (2013). Network analysis of third party tracking: User exposure to tracking cookies through search. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 549–556.

Pew Research Center (2014). State of the news media 2014: Key indicators in media and news. Technical report.

PricewaterhouseCoopers (2015). IAB internet advertising revenue report. Technical report, Interactive Advertising Bureau.

Reisman, D., Englehardt, S., Eubank, C., Zimmerman, P., and Narayanan, A. (2015). Cookies that give you away: Evaluating the surveillance implications of web tracking. In *Proceedings of the 24th International World Wide Web Conference*. International World Wide Web Conference.

Roesner, F., Kohno, T., and Wetherall, D. (2012). Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 12–12. USENIX Association.

Schmeiser, S. (2015). The size distribution of websites. *Economics Letters*, 128:62–68.

Schmeiser, S. (2016). Sharing audience data: Strategic participation in behavioral advertising networks. *Working Paper*.

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*, pages 261–270. ACM.

# A Data details

## A.1 Special interest categories

The following subcategories are classified as special interest sites. Each aligns with a particular advertising market.

Table 8:

| Subcategory | No. sites |
|---|---|
| Alternative and Natural Medicine | 4 |
| Baseball | 7 |
| Basketball | 3 |
| Board and Card Games | 4 |
| Child Health | 4 |
| Cycling and Biking | 8 |
| Fantasy Sports | 7 |
| Flowers | 4 |
| Football | 15 |
| Furniture | 14 |
| Gardening | 2 |
| Genealogy | 9 |
| Golf | 7 |
| Home Improvement | 12 |
| Immigration and Visas | 4 |
| Interior Decor | 4 |
| Martial Arts | 7 |
| Mental Health | 3 |
| Motorcycles | 4 |
| Motorsports | 4 |
| Nursery and Playroom | 12 |
| Nutrition | 7 |
| Outdoors | 4 |
| Pets | 15 |
| Roleplaying | 7 |
| Running | 7 |
| Soccer | 10 |
| Tennis | 3 |
| Theme Parks | 4 |
| Water Sports | 2 |
| Weddings | 10 |
| Weight Loss | 1 |
| Winter Sports | 3 |
| Womens Interests | 14 |
| Total | 225 |

## B Top 100 sites

Table 9 replicates Table 4 but for the top 100 sites rather than the top 10,000.

Table 9: Statistics on the number of networks per site for the top 100 sites. The first column "Sites" reports the number of sites in the category. "Share adnet" reports the share of sites in the category that use one or more advertising networks. "Sites adnet" reports the number of sites in the category that have one or more advertising networks. The next four columns report the minimum, median, mean, and max number of ad networks per site among sites that have one or more advertising networks. "Share OBA" reports the share of sites (conditional on using advertising networks) that use one or more behavioral networks. The next columns report the minimum, median, mean, and maximum number of behavioral networks per site conditional on the site using advertising networks.

| | | Share | Sites | Ad networks | | | | Share | Behavioral networks | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Sites | adnet | adnet | Min | Median | Mean | Max | OBA | Min | Median | Mean | Max |
| Adult | 2 | 1 | 2 | 1 | 1.500 | 1.500 | 2 | 0 | 0 | 0 | 0 | 0 |
| Arts and Entertainment | 8 | 0.875 | 7 | 1 | 3 | 3.571 | 9 | 0.857 | 0 | 2 | 1.857 | 4 |
| Business and Industry | 5 | 1 | 5 | 1 | 3 | 3.400 | 9 | 0.800 | 0 | 1 | 1.800 | 5 |
| Career and Education | 2 | 1 | 2 | 2 | 2.500 | 2.500 | 3 | 0.500 | 0 | 0.500 | 0.500 | 1 |
| Computer and Electronics | 7 | 0.571 | 4 | 1 | 2.500 | 3.750 | 9 | 0.750 | 0 | 1 | 1.750 | 5 |
| Finance | 5 | 1 | 5 | 1 | 3 | 3.400 | 7 | 0.800 | 0 | 2 | 2 | 4 |
| Games | 1 | 1 | 1 | 5 | 5 | 5 | 5 | 1 | 3 | 3 | 3 | 3 |
| Health | 2 | 0.500 | 1 | 9 | 9 | 9 | 9 | 1 | 4 | 4 | 4 | 4 |
| Internet and Telecom | 22 | 0.455 | 10 | 1 | 3.500 | 4 | 9 | 0.800 | 0 | 2 | 1.800 | 4 |
| Law and Government | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| News and Media | 19 | 1 | 19 | 2 | 5 | 5.895 | 11 | 1 | 1 | 3 | 2.947 | 5 |
| People and Society | 1 | 1 | 1 | 12 | 12 | 12 | 12 | 1 | 5 | 5 | 5 | 5 |
| Reference | 2 | 0.500 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| Shopping | 9 | 0.889 | 8 | 1 | 2 | 2.625 | 8 | 0.625 | 0 | 1 | 1.500 | 6 |
| Travel | 1 | 1 | 1 | 9 | 9 | 9 | 9 | 1 | 6 | 6 | 6 | 6 |
| Uncategorized | 13 | 0.769 | 10 | 1 | 2.500 | 3.200 | 6 | 0.900 | 0 | 1.500 | 1.600 | 3 |
| All Sites | 100 | 0.780 | 78 | 1 | 3 | 4.167 | 12 | 0.833 | 0 | 2 | 2.077 | 6 |