The Internet as Recommendation Engine: Implications of Online Behavioral Targeting

by

Anthony N. Smith-Grieco

B.A., Reed College (1996)

Submitted to the Engineering Systems Division
in partial fulfillment of the requirements for the degree of

Master of Science in Technology and Policy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2010

Author . . . .
Engineering Systems Division
Jan 15, 2010

Certified by . . . .
David D. Clark
Senior Research Scientist
Computer Science and Artificial Intelligence Laboratory
Thesis Supervisor

Accepted by. . . . .
Dava J. Newman
Professor of Aeronautics and Astronautics and Engineering Systems
Director, Technology and Policy Program

The Internet as Recommendation Engine: Implications of Online Behavioral Targeting

by
Anthony N. Smith-Grieco

Submitted to the Engineering Systems Division
on Jan 15, 2010, in partial fulfillment of the
requirements for the degree of
Master of Science in Technology and Policy

Abstract

This thesis discusses the economic implications of Internet behavioral advertising, which targets ads to individuals based on extensive detailed data about the specific websites users have visited. Previous literature on behavioral advertising has focused almost exclusively on privacy issues; there has been less study of how it might affect industry structure. This thesis examines which parties in the online advertising value chain would benefit the most from the demand for detailed behavioral data; in particular, it examines whether aggregators (such as advertising networks) that track behavior across a large number of websites would derive the greatest benefit. Qualitative stakeholder analysis is used to identify the strengths and weaknesses of several categories of actors: advertisers, advertising agencies, publishers, advertising networks, advertising exchanges, Internet service providers, and users. Advertising agencies might attempt to bypass networks and work directly with publishers, becoming aggregators in their own right. Publishers might need to become interactive “information experiences” in order to collect valuable behavioral data. Users might demand more transparency about what is happening with their data, or even more control over the data collection process. Overall, agencies, networks, and advertising exchanges appear to be in the best position; publishers are faced with a harder task. Furthermore, behavioral targeting may not result in a dramatic increase in overall online advertising spending.

Thesis Supervisor: David D. Clark
Title: Senior Research Scientist
Computer Science and Artificial Intelligence Laboratory

Acknowledgments

First of all, I must express my immense gratitude to my advisor, David Clark, for many things, starting with his willingness to take me on as a student, and share his extensive knowledge and experience with me. He was exceedingly patient with me as I worked to define a topic and then develop it into a finished product. He gave much of his valuable time, and read my drafts amazingly quickly. In short, he has made this process as easy as it could be. In addition, he helped me to go to several conferences, for which I am very grateful.

Thanks to my officemates, Jesse Sowell, Steve Woodrow, and Steve Bauer, for their advice and feedback on my ideas, and for many fun conversations. Special thanks to Chintan Vaishnav for sharing his experience, wisdom and very kind heart, and for reading an earlier draft and offering his feedback. Susan Perez has continually been helpful with the little things, and made the CSAIL ANA office a warm place to work.

Thanks to all of the people in the online marketing field that I spoke with in the course of my research, for providing valuable inside perspectives; especially, Kate MacKenzie, Avinash Gupta, Amy Kang, and Dave Donars.

Thanks to Fernando Bermejo from the Berkman Center for taking an interest in my work. Dan Pereira from the Convergence Culture Consortium has been especially encouraging and pointed me towards new ways of thinking about the issues. Thanks to Subu Desaraju for sharing his inside perspective and allowing me to bounce ideas off of him. Thanks to Frank Field for guidance in defining my topic and for a key reference.

Thanks to Bill Lehr, Nazli Choucri and the MIT Joint Program on the Science and Policy of Global Change for the opportunity to work as a research assistant on their projects. I would especially like to thank Sebastian Rausch, my other supervisor, for being very accommodating while I was in the last phase of the thesis process.

Thanks to my parents, Sid and Greg, for their love and support, and my mother in particular for setting aside many hours to work with me, and offering her advice and encouragement when I needed it.

And finally, the biggest thanks of all goes to my wife Claudia, for her support, patience and love during this process, especially when it consumed much more of our life together than was reasonable.

Biographical Note

This thesis was undoubtedly shaped by my years of experience as a software developer. In particular, I worked for Yahoo, one of the companies discussed in this thesis, in 2005 and 2006, although I did not work on an advertising product. Earlier in my career, I also deployed and managed an advertising system for several publisher websites. The opinions offered in this thesis reflect my own views, and not those of my former employers.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

Imagine that upon signing on the Internet, the ads one sees are a reflection of the websites he has visited and the products he has read about. Every website one visits is factored into an engine which is also observing everybody else's browsing patterns and the products they take an interest in. The engine assumes that if Roger's browsing patterns are similar to John's, then maybe he will buy the same things as John – so why not show the same ads. Because Roger looked at a review site for digital cameras, he starts seeing ads about digital cameras even when reading the news. If he reads blogs about parenting, he'll start to see ads for baby products.

In short, the entire Internet would become a *recommendation engine*, similar to what is seen on Amazon.com and other online retailers: based on the products you have looked at, Amazon suggests others that you might find of interest[1]. The difference is that while Amazon prefaces the recommendations with "people who viewed these products also looked at…", the reasoning behind these ads will be invisible. In addition, they will not purely be based on your past browsing, but also which advertisers are paying more. And finally, the value chain for this recommendation engine is much more complex, as it incorporates a range of stakeholders: individual publishers, advertisers and intermediaries, such as ad networks.

This is the future on which some in the digital marketing world are betting. The technological architectures of online audience measurement enable sophisticated data mining and analysis of user browsing habits that increase the economic value of advertising to marketers and publishers precisely because users can be segmented into increasingly customized categories. Drawing on the heritage of direct marketing, online marketers are attempting to make online advertising more finely targeted to individuals based on knowledge about what websites individuals have visited (Turow, 2006). Looking at these expressions of interest, advertisers target car ads to those who appear to be shopping for cars, and show offers for discounted flights to those who have been shopping for flights. At the same time, newspapers and other media outlets are seeing advertising revenues decline as people spend more time on the Internet. Online advertising, though growing, is not yet as big a revenue source as print and radio and TV. This is challenging media outlets to consider new business models or find ways to increase the value of their advertising inventory, potentially by making it more targeted (Downie Jr. & Schudson, 2009).

This kind of targeting is based on collection of data by parties called *aggregators*, here defined as parties that collect data from a wide range of websites and/or users by virtue of their extensive business relationships with those websites' publishers. Advertising networks, including Google and Yahoo, are the most obvious aggregators, but it is also possible for other parties, such as advertising agencies, to play that role. This thesis examines whether the new measurement architectures that are developing are likely to privilege large data aggregators relative to

[1] The connection between behavioral advertising and recommendation engines is noted by two Yahoo researchers in the materials for a Stanford course on "computational advertising" (Broder & Josifovski, 2009).

individual web publishers. In the same way that credit bureaus, with their knowledge of individuals' financial histories, have great power to determine individuals' access to credit, perhaps these data stores will come to decide which ads people will see. Furthermore, as marketing becomes more and more about data, a battle will develop over who has the data and who can use it. Small publishers, unable to generate the volume of data needed for these tracking engines, might find themselves at a disadvantage. The large advertising agencies might try to bypass the intermediaries they currently work with, in order to have more direct access to the data. Users could themselves start claiming a stake in the "value" they produce with their online behavior.

The question of market power is not new to this space; it was raised when Google acquired DoubleClick, one of the leading advertising networks. At the time, Google was the leader in search advertising, whereas DoubleClick was a prominent player (but not dominant) in display advertising. Neither party was competing in each other's business directly at the time; Google did not do display ads, and DoubleClick did not do search ads. The FTC evaluated several possible ways in which a merger might threaten competition, but in the end allowed the merger, ruling that the two markets in which the companies operated were separate enough (US Federal Trade Commission, 2007). Notably, it did not consider the privacy implications of such a merger, holding that privacy was a separate concern to be evaluated by a separate department of the agency. However, we now begin to wonder if such a benign interpretation can be sustained, and in fact the Department of Justice is believed to be considering antitrust action against Google[2].

Much of the focus on behavioral advertising has been on privacy issues, for good reason. Congressman Rick Boucher has introduced legislation to address privacy issues with online advertising, potentially limiting how data is collected or used. Even if marketing is deemed to be an acceptable motive for collection of personal information, there remains that chance that such information might also be used for surveillance, policing or anti-terrorism purposes, if governments were to pressure private companies to share such information. However, this thesis will not grapple directly with the privacy issues raised by behavioral tracking, which have been the subject of extensive discussion from a variety of points of view, including philosophy, law, sociology, and computer science. The goal here, rather, is to better understand the economic consequences, and to suggest ways that policymakers might think about the economic implications of privacy regulations. This is not to say that the privacy issues are unimportant, but simply that there are other dimensions to the topic[3].

This paper is informed by a reading of the trade press and discussions with several industry actors. However, it makes no pretense of following an exhaustive, rigorous methodology. Economic reasoning and concepts will be used, but no formal economic model will be developed. Limited quantitative data is presented, but this data is not sufficient for a quantitative study. In some ways, the thesis takes the form of a stakeholder analysis, starting from an understanding of the goals of each of the actors in this system and the means they might pursue to achieve those goals. That knowledge is then used to conduct a kind of "thought experiment"

[2] "Return of the Trustbusters" (August 27, 2009), *The Economist*

[3] See also Baker (2008) for a discussion of how behavioral data mining technology might change other areas of life, such as politics and the workplace.

about what might happen if individual profiles were to be come highly valuable property in the marketing ecosystem. This is not to say that other kinds of assets, such as technology, institutional relationships between advertisers and media channels, and the content of advertising messages, are not also important. The purpose of this thesis is to start from an assumption, that data about individuals is economically important, and explore the implications of that assumption, in order to produce some testable hypotheses. If certain kinds of behavior are observed in the market, there would be added reason to conclude that the assumption is correct. On the other hand, if the implications described in this thesis do not come true, then the initial assumption would be shown to be false. This thesis engages in a kind of analysis that might be forced upon a policymaker with a need to produce some judgments but limited access to information about a fast-moving industry. It is also influenced by earlier studies of how network architecture shapes the balance of power among the Internet's stakeholders, and how those stakeholders can in turn attempt to restructure the network within limited parameters (Blumenthal & Clark, 2001; Zittrain, 2006).

One important point sometimes lost in the discussion of behavioral advertising is that the observation of users' behavior, and the targeting of ads based on that behavior, are two separate things. Observation of users can occur even when no ads are displayed; and likewise, an ad can be displayed without recording user behavior. The observer or collector of data need not be the same party that places the ad based on the data. Thus there are actually two separate (but related) markets at work: the market for user data, and the market for online advertising. Prices in the two markets may be connected, but in ways that are difficult to understand.

In this thesis, the terms "advertiser" and "marketer" will be used somewhat interchangeably, although that is perhaps not correct in the strictest sense. For some, "marketing" encompasses a broad spectrum of activities beyond simply advertising, such as market research, promotions, customer-relationship management and public relations. However, for this thesis, the focus is specifically on advertising, which is defined as the purchase of space and/or time in a media channel for the purpose of displaying a message about one's product. Yet, the Internet enables and encourages individuals to actively choose how they engage with media content, and even allows them to avoid advertising altogether if they are savvy enough (by means of add-ons to Internet browsers). The Internet also creates possibilities for new kinds of marketing activities, based not around the "broadcast" of a message or image to a huge number of people, but attuned more to the conversational nature of the medium. Marketers no longer have total control over their brand messaging. It used to be that an advertisement would be created and consumers would passively receive it via TV or radio, with no opportunity to talk back or respond. Now, however, consumers can post reviews of products, and create and share their own videos about products on YouTube. Spurgeon (2008) gives the following example:

"Home videos of explosive Coke-Mentos soda fountains and Coke-Mentos rockets started appearing on the Web in early 2006. This association of Coke with a lesser brand of mints took both brands by surprise. The brand companies could control neither the uses made of their products, nor the dissemination of the images of those uses. The replication, video capture, and Web-based sharing of Coke-Mentos experiments snowballed... Mentos was very happy with this popular appropriation and display of its brand, and its association with youth culture values. It estimated this media exposure was worth US $10 million, equivalent to more than half its annual

advertising budget for the US market (Vranica and Terhune 2006), and took immediate steps to build on this publicity opportunity by partnering with YouTube to run a competition for the best Coke-Mentos video. Although early responses from Coke were not enthusiastic, the global soft drink giant also elected to explore this consumer-generated media activity as a brand-building opportunity. It mounted a 'Poetry in Motion' competition that challenged Coke consumers to show the world what extraordinary things they could do with everyday objects (Vranica and Terhune 2006)." (p. 1)

It is possible that marketing will become as much about *listening* to what people are saying online about your brand, as it is about speaking to them *en masse* (Wetpaint & Altimeter Group, 2009). In addition, marketers can attempt to focus their efforts on small groups of influential bloggers, and use them to spread the word about products. This kind of activity has gained the attention of the FTC[4]. Or, marketers can create small Facebook applications that people share with their friends. Jenkins (2006) argues that marketers are increasingly looking to form more of an engagement with consumers. Simply viewing an ad is no longer enough; marketers want consumers to express their identification with a brand by for example sharing personal stories about what the product means to them. If that is true, there may be a demise of "traditional" online advertising models focused on pushing messages at consumers, in favor of models focused more on interaction and engagement. However, this thesis will not examine that possibility in depth; the focus here is more on how traditional advertising practices have adapted to the Internet.

The paper proceeds as follows. Chapter 2 gives an overview of the Internet advertising ecosystem, describing the roles played and value offered by advertisers, publishers, advertising networks, and Internet service providers. Chapter 3 explores several important topics in more depth. It explains how behavioral advertising differs from other kinds of ad targeting, gives some indication of the monetary value associated with behavioral advertising, and also briefly discusses online audience measurement, as another kind of data collection that can be compared with behavioral tracking. Chapter 4 will explain the technology behind the tracking of users, the tools available to avoid such tracking, the kinds of "identities" attached to users, and the data available to various kinds of actors. Finally, chapter 5 will return to our hypothesis and examine the ways in which market power might arise, and how various actors could challenge the concentration of data in the hands of aggregators.

[4] Stephanie Clifford, "Notice Those Ads on Blogs? Regulators Do, Too" (August 10, 2009), *New York Times*

Chapter 2. The Internet advertising ecosystem

2.1 Introduction

This chapter will briefly introduce some of the key players in the online advertising ecosystem, advertising agencies and publishers. It will also discuss how various kinds of intermediaries, the advertising networks and advertising exchanges, have arisen to connect these two parties.

Internet advertising is a market where websites sell space on their webpages to advertisers, who buy this space because they have a message to convey to the website's audience. Websites refer to this space on their pages as their *inventory*. Online ads can be further subdivided into several categories: search ads, display/banner ads, video ads, and mobile ads (ads on mobile phones). This paper is focused largely on banner ads and video ads. Banner ads were the first form of online advertising; a visual graphic appearing somewhere on the page, which when clicked, led to the advertiser's site. Video advertising is a new, early-stage market, just developing now as the consumer appetite for online video has increased. The advertising itself may be a video played before or after the content, or an overlay on the top, bottom or side of the video window.

Table 1. US advertising and marketing spending, online and offline, 2008 (billions of $$)[5]

Internet	**23.4**
Search	10.5
Display and rich media	6.5
Classifieds	3.2
Lead generation	1.7
Video	0.7
Email	0.4
Sponsorships	0.4
Offline advertising	
TV	57.9
Newspapers	44.0
Magazines	23.6
Radio	19.2
Other marketing services	
Sales promotion	76.4
Telemarketing	47.0
Direct mail	49.1
Event sponsorship	21.2
Directories	13.2

As illustrated in the attached table, search ads are estimated to represent about a half of all online advertising spending; display and rich media (which includes interactive forms of visual ads) account for slightly less than search. It is also worth noting that the combined amount spent on all forms of online advertising is still appreciably less than the amount spent on TV advertising, newspaper advertising and various kinds of "marketing services", including direct mail.

Lead generation, as defined by the Interactive Advertising Bureau, refers to "fees advertisers pay to Internet advertising companies that refer qualified purchase inquiries (e.g., auto dealers which pay a fee in exchange for receiving a qualified purchase inquiry online) or provide consumer information (demographic, contact, behavioral) where the consumer opts into being contacted by a marketer (email, postal, telephone, fax). These processes are priced on a performance basis (e.g., cost-per-action, -lead or -inquiry), and can include user applications (e.g., for a credit card), surveys, contests (e.g., sweepstakes) or registrations" (Interactive Advertising Bureau & PricewaterhouseCoopers, 2009).

Advertising is commonly divided into two types: *branding* and *direct response*. Brand advertisements have the goal of simply promoting awareness of their brand; they are not necessarily expected to elicit a purchase right away. A brand advertisement may be based more around telling a story designed to raise interest or create desire in a large, mass audience. Direct

[5] For Internet advertising: eMarketer, "US Online Advertising Spending, by Format, 2008-2013", accessed August 19, 2009. For non-Internet advertising: eMarketer, "US Advertising and Marketing Spending, by, Media, 2008-2011", accessed November 22, 2009; original source is Zenith Optimedia.

response advertisements, on the other hand, have some immediate expectation of interaction or engagement with the user: for example, “Call this number to receive a discount”. The distinction between branding and direct response is not hard and fast; a 30-second TV advertisement can tell a story but then invite the viewer to visit the company’s website for more information. Traditional brand advertising has been slow to migrate online; the majority of online advertising falls in the direct response category (Nielsen, 2009b).

2.2 Advertisers and ad agencies

Traditionally, all but the smallest advertisers have relied on advertising agencies to do much of the work of designing ad campaigns and buying ad space. Agencies can offer a variety of services, ranging from creative development (producing the audio, video, or graphical elements of advertisements) to media buying (negotiating agreements to place ads with TV stations, newspapers, or other media outlets), media planning, public relations, and direct response marketing. Agencies used to be relatively small organizations, but over the past few decades, there has been tremendous consolidation in the agency business, with the result that there are now a small number of holding companies which control almost all of the major agencies: WPP, Omnicom, Publicis, and IPG. To some degree, these holding companies offer their clients (the advertisers) “one-stop shopping” for all of the different kinds of services listed above. For example, the holding company Interpublic Group (IPG) owns the agencies Campbell-Ewald, Hill Holliday, McCann Erickson, Mullen, and Rogers & Cowan (just to name a few). These individual agencies retain their own identity, despite their integration into the holding company[6].

It used to be that agencies would be paid on commission for media buys, i.e. they would take 15% of the amount paid for a TV, radio, or other advertising spot (Cappo, 2003). This compensation model incentivized the agencies towards big purchases of time and/or space in major media, and discourages small, incremental, experimental online campaigns where the cost of buying media (ad inventory) is relatively low. In addition, agency clients (the advertisers) are pushing for more accountability from marketing spending, with measurements of performance and results. Thus there is a trend towards a fee-for-service model, potentially time-based and with performance incentives[7]. Measuring the performance of a direct response campaign is relatively straightforward: ad clicks, site registrations and online purchases can all be quantified. Measuring the performance of a brand campaign, on the other hand, is more difficult, and may involve brief surveys of people, or more complicated longer-term statistical analysis of consumption patterns.

In addition to the pressure for measurement and continual refinement, in the online environment large agencies find themselves in an ambiguous relationship with technology companies like Google, Microsoft, and Yahoo. Agencies may be feeling some pressure from technology companies and advertising networks, as advertising becomes increasingly data-driven and quantitative: both the ad agencies as well as the technology companies are in the business of analyzing consumer behavior in order to predict ad effectiveness. Potentially the technology

[6] The TV series *Mad Men* offers an interesting portrait of an agency in the early 1960s, before the rise of the holding companies.

[7] “Clock-watchers no more”, (May 14, 2009), *The Economist*

companies like Google could try to "disintermediate" the agencies out of business. The ad agency holding company WPP devoted a small section of its 2008 annual report to the question of whether Google is a friend or a foe (WPP, 2009a). WPP forecasted spending $850 million on Google search advertising in 2009, making it Google's largest agency customer. Google and WPP also are joint sponsors of a three-year, $5 million research program on the effectiveness of online advertising, overseen by faculty from Harvard Business School and MIT/Sloan (WPP, 2009b). In its annual report, WPP reviewed the complex relationship it has with Google, and offered these comments:

"All in all, Google is opening up the attack on many fronts. Perhaps too many, particularly when you consider the other theatres it is fighting in, such as book publishing and robots to the moon. One gets the impression it is throwing a lot of mud against the wall to see if any sticks - maybe sticking to mobile search would be best. Yahoo! has a different approach, working through its agency partners and believing in the power of people, rather than Google's greater focus and belief in technology. Certainly, even now, a combination of Microsoft and Yahoo! in any way will bring greater balance to the markets. Our clients and our agencies will favour a duopoly rather than a monopoly."

2.3 Publishers

Publishers are individuals or companies that produce websites, and are dependent on advertising revenue. The terminology here is a bit awkward, because within the online advertising industry, it is customary to refer to any party that sells ad inventory as a "publisher", even websites that are not authors or producers of their own content. For example, an online retailer could be considered a "publisher" to the extent that it also makes revenue by selling ad inventory. Nonetheless, advertising revenue is more of a concern for the classic kind of publisher whose entire focus is the production of content and the sale of advertising inventory alongside the content, such as newspapers, magazines and blogs. As newspapers and magazines have migrated from entirely print media to mixed print and online forms, their ad revenue has decreased; spending on online advertising has not compensated for a decline in spending on traditional print and TV advertising (Downie Jr. & Schudson, 2009).

In the online environment, as with other media, publishers' challenge is to compete for consumers and to demonstrate that their audiences have value for advertisers. In essence, publishers are selling audiences. As different advertisers will be interested in different audiences, the challenge for publishers is to understand their audiences in such a way as to be able to sell their sites to advertisers.

Publishers are also concerned about showing ads on their sites that alienate the users who are invested in the information, ideas, and imagination of the publisher. The key question for publishers is simply how to make the most advertising revenue from the site, without irritating users by overwhelming them with ads. The trade-off is between placing more ads on the page, which might increase revenue in the short term, but might push away or offend users in the longer term. Some argue that publishers need to create a scarcity of advertising space by intentionally reducing the number of advertising spots on a page.

Publishers generally divide their inventory into two categories: “premium” and “remnant”. Premium inventory is the inventory that can be sold directly to advertisers at a high price; for example, the ad spots on the home page of a respected news site which may be seen by millions of people in the course of a day. Remnant inventory is whatever inventory can not be sold directly to advertisers, perhaps because it is on obscure pages that are visited less often or do not have content with which advertisers prefer to be associated. Of course, this distinction is fluid, and remnant inventory could become premium if packaged in the right way to advertisers (Winterberry Group, 2009).

2.4 Advertising networks

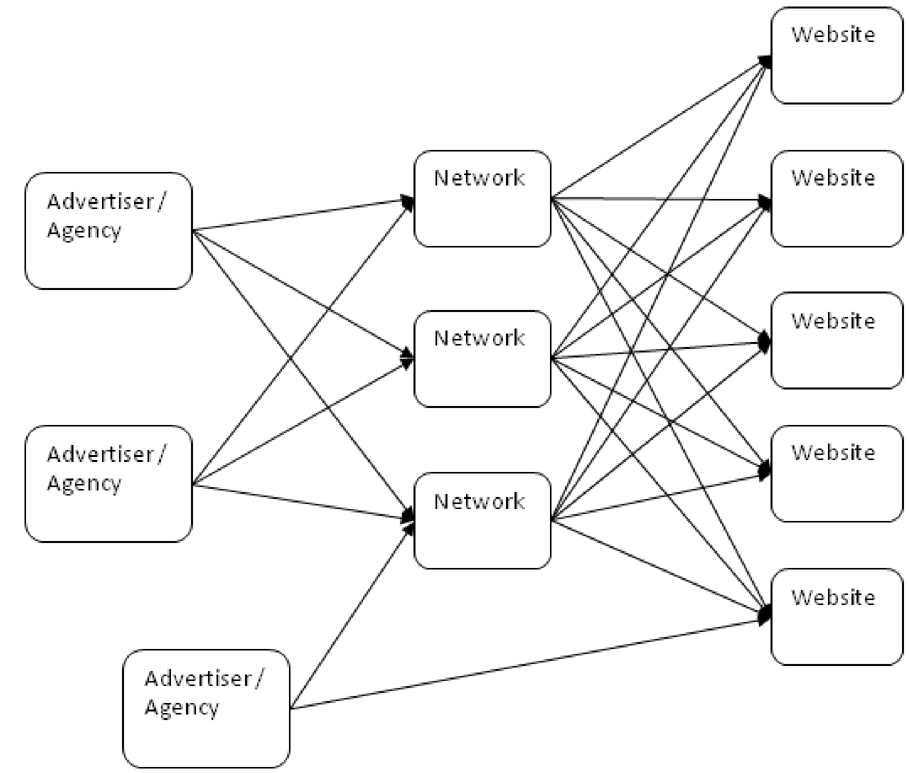
Popular, well-known websites and big-spending advertisers make deals directly with each other, but the remainder of the market works through intermediaries such as ad networks. Ad networks have several functions in the advertising ecosystem:

- enabling scale / reducing transaction costs by making it possible to advertise on a wide variety of sites without having to make deals directly with a great number of publishers
- for publishers, screening for ads which may not be appropriate for the publisher’s site (for example, because they link to adult content or disreputable advertisers)
- for advertisers, a kind of risk management: the advertiser may pay the network only when someone clicks on the ad, so impressions that yield no response do not lose money
- collection and analysis of individuals’ browsing patterns and responses to different kinds of ad campaigns
- related to the previous point, the ability to target ads to individuals based on their browsing habits or other variables (such as location and demographics)

Advertising through an ad network works as follows: first, an advertiser defines an *ad campaign*, which is a set of parameters that describe where and when ads should appear. For example, a campaign might specify that a particular ad should be displayed two million times on sports-related websites for users coming from the West Coast during the second week of February. Based on these parameters, the ad network quotes a price for the campaign, generally based either on the number of ad displays (also referred to as *impressions*) or the number of clicks on the ad. The ad network then is responsible for deciding which pages on which sites will show the ad. It pays the publishers of those sites at a lower rate than the advertiser is paying, and pockets the difference as its own margin.

The attached figure illustrates how money might flow between advertisers, networks and websites. An advertiser might buy ad placements through several networks at the same time, and each of those networks may in turn place ads on several websites. The advertiser at the bottom buys ad placements both through networks and directly from websites. A website may receive payments and place ads from a number of networks, and also directly from advertisers. In short, it is generally not a simple “one-to-one” relationship between buyer and seller, but rather a “many-to-many” relationship.

Figure 1. Money flow between advertisers, agencies, advertising networks and websites



(The arrows point from the buyer to the seller. In other words, advertisers pay networks, which in turn pay websites.)

Direct sales are generally more lucrative for publishers, and possibly more effective for advertisers, or at least more customizable. They get more control over when, how and to whom the ads are shown. On the other hand, ad networks are generally cheaper for advertisers, and provide a cost-effective way to reach a large audience quickly and easily. For publishers, networks are a way to gain additional revenue from their ad inventory above what can be gained through direct sales, generally from remnant inventory that could not be sold directly. Smaller publishers, lacking the name recognition to sell directly to big advertisers, may rely exclusively on networks for their revenue.

A recent survey of advertisers[8] reported that 90% of advertisers plan to work with ad networks in

[8] eMarketer, "Who Loves Ad Networks?" (May 18, 2009), http://totalaccess.emarketer.com.libproxy.mit.edu/Article.aspx?R=1007091, accessed July 20, 2009

2009. But the majority of those advertisers spend 15% or less of their budget on networks.

According to some estimates there are several hundred ad networks in existence today. There is even a company Adify which develops the technological infrastructure for developing a network, which suggests that networks are becoming commoditized from a technological point of view. The market research firm comScore releases public rankings of ad networks based on their "reach" -- a measurement of the number of unique users that see ads from these networks. This measure does not speak to the number of ads delivered by the networks, nor to the monetary value of those ads, but only to the size of the audience accessible to the networks. The attached table shows this data for April 2008 and April 2009, illustrating how much the market changes in the course of one year. The TV network Fox has a new Internet advertising network. Several other networks showed dramatic jumps from the previous year: 24/7 Real Media, Turn, CPX Interactive. A number of other networks also showed some growth. At a minimum, this data suggesets that Internet users may be seeing ads from a much greater variety of networks than before.

Table 2. Top 25 US Ad Networks, by Reach (in April 2009), from comScore[9]

Ad Network	Total Unique Visitors (thousands)		
	April 2008	April 2009	% change
Total Internet Audience	190,728	192,875	1
AOL Platform-A	170,508	176,455	3
Yahoo! Network	160,206	167,129	4
Google Ad Network	155,882	164,518	6
ValueClick Networks	140,930	160,307	14
Specific Media	144,773	158,012	9
FOX Audience Network	N/A	149,249	N/A
24/7 Real Media (WPP)	99,959	147,668	48
Traffic Marketplace	114,682	143,519	25
Microsoft Media Network US	119,595	139,674	17
Tribal Fusion	135,113	138,274	2
Casale Media - MediaNet	127,184	137,884	8
interCLICK	107,961	134,834	25
Turn, Inc	60,617	134,028	121
Adconion Media Group	117,965	133,498	13
CPX Interactive	69,178	130,370	88
Collective Network by Collective Media	88,279	129,808	47
ADSDAQ by ContextWeb	93,815	123,534	32
AudienceScience (formerly Revenue Science)	N/A	121,001	N/A
Burst Media	89,670	116,727	30
Undertone Networks	72,940	97,053	33
AdBrite	81,838	91,033	11
Pulse 360	N/A	82,574	N/A
Vibrant Media	72,351	80,779	12
Adify	N/A	73,467	N/A
Kontera	52,159	72,870	40

This plethora of networks may be classified along several dimensions: the sites (publishers) that are part of the network, the level of transparency to advertisers, and whether or not they are owned by a larger media or technology company (Advertising Age, 2009; Karpinski, 2009a; Karpinski, 2009b).

Sites targeted: Some networks work only with “premium” publishers that guarantee a “safe”, well-known, trusted environment for their advertisers. Other networks focus more on the “mid-tail” and “long tail” of sites. The “long tail” is a term of art for the vast number of niche websites which individually have small audiences (perhaps in the thousands or at most hundreds of thousands of people). Advertisers may view these sites as more “risky” in the sense that they know less about the content their ads will appear against, but those sites generally sell their

[9] Source: comScore press release, “comScore Releases April 2009 U.S. Ranking of Top 25 Ad Networks” (May 20, 2009), http://www.comscore.com/Press_Events/Press_Releases/2009/5/Top_25_US_Ad_Networks, accessed July 20, 2009

inventory at lower rates. Also, there are so-called “vertical” networks focused on a specific industry or topic area, such as health care, politics, or sports. For example, the MTV Tribes network is focused on the youth, music and entertainment vertical.

The same site might sell ad space via a number of networks, but at a different price for each. For example, a mid-tail political news site might sell ad space through an general ad network as well as a politics-specific vertical network, but the latter might pay a higher rate because it is presenting the site as part of a more valuable “bucket” of sites of similar kinds. An advertiser interested in reaching a politically-savvy audience might be willing to pay more to a politics-oriented network than to a general network (even if the latter network offers a “politics” channel).

Transparency: This refers to the question of how networks balance advertisers' need to know something about where ads are being placed, with the network's need for control over this “proprietary” information. If a network revealed every site on which an ad was displayed, the advertiser would know that it can place an ad on, say, the *New York Times* site much more cheaply through the network than through a direct deal with the publisher. On the other hand, advertisers do want to know something about where their ads are displayed. To address this need, the network might reveal categorical information: for example, the sites are in X and Y categories of sites, or they are in the top 1,000 web sites in terms of traffic.

Even if the network does not share this data, advertisers and publishers may learn something about where ads are being shown from independent market research firms such as comScore and Nielsen. However, as will be discussed later, this data also has limitations as it is usually derived from a sampling of Internet users.

In some cases, advertisers are concerned about their ads showing up in objectionable places, such as adult sites, or competitors' sites. Thus some networks allow advertisers to specify a “blacklist” of forbidden sites where an ad should not appear.

Ownership: Some networks are started by brand-name publishers, as a way of increasing the volume of ads they can sell by incorporating other publishers into their orbit. For example, MTV started an advertising network called Tribes which includes other sites that MTV feels are a good fit for its brand image. NBC and Fox have also formed such networks. In a sense, publisher-affiliated networks are a kind of outsourced ad sales team. On the other end of the spectrum are networks affiliated with ad agencies, which may be seen as ways for agencies to coordinate ad buys across all of their accounts, as well as to “cut out the middleman” -- i.e., the other networks. 24/7 RealMedia is one example; initially it was a large, independent network, but was later acquired by the ad agency holding company WPP. Then in the middle of this spectrum are “independent networks” which are not tightly affiliated with any single publisher brand.

Other networks are owned by larger media or technology conglomerates, such as Time Warner/AOL, Yahoo, Google, or Microsoft. AOL, for example, has recently made a big effort to enter the online advertising space, with a number of purchases of ad networks. When a network is owned by a company with a content division (such as AOL), there may be an incentive to privilege the in-house publishers over outside publishers and sites[10].

One investment banking report (DeSilva + Phillips, 2008) lists about 15 ad network acquistions in the year 2007 alone, as well as several more acquisitions in previous years. AOL purchased several of these, and Microsoft, Yahoo and Google were responsible for a couple each. WPP purchased 24/7 Real Media for about $650 million. The largest acquisition was of the aQuantive (DrivePM) network by Microsoft, for about $6 billion. Google's acquisition of DoubleClick was second, $3.1 billion. The remaining acquisitions were on the order of tens or hundreds of millions of dollars.

The variety of networks, acquisitions and ownership patterns suggests that there might be several different kinds of *economies of scale* at work in this market. Vertical ad networks indicate that there might be an economy of scale in knowledge of user behavior and interests in particular topic areas (such as travel, sports or real estate). The more broadly-focused networks suggest that there might be an economy of scale in knowledge of user behavior across multiple areas – in other words, there is value in developing a broad, multi-faceted portrait of the potential consumer from their behaviors across many kinds of sites. For example, perhaps knowing about users' financial activity might be helpful for marketing travel products. If this were the case, then portals like AOL and Yahoo would also benefit, because of their variety of different kinds of content and places to interact with users. Finally, there may be economies of scale simply with regards to the number of users reached by the network, and/or the number of advertisers buying from the network. These different economies of scale would have different implications for the kinds of market power that might develop.

One final point to note is that advertising networks are not really "networks" in the same sense as social networks like Facebook. There is no concept of "friendship" in an ad network, and no complex "social graph" that one might wish to study or traverse. It is interesting to contemplate if a new kind of ad network could be developed based on affinity relationships; for example, groups of advertisers could indicate their connectedness, and thus be targeted in similar ways, and likewise groups of publishers might indicate their relatedness and thus receive similar ads. Even individuals could express their preferences for particular brands or advertisers, and then receive offers from others in those advertisers' affinity networks. In some ways, this is already happening on Facebook, where individuals can indicate that they are "fans" of a particular brand or company.

2.5 Ad exchanges

As ad networks have proliferated, there has been some concern among publishers and advertisers both about the general transparency and efficiency of the market, given that multiple networks might place the same ad on the same site for different prices. This has given rise to *ad exchanges,* platforms where networks, publishers and advertisers can buy and sell inventory in an auction framework. ContextWeb's ADSDAQ, Yahoo's Right Media and APT, and Google's DoubleClick exchange are the most prominent exchanges. Generally networks and agencies bid

[10] David Koretz, "Ad Networks Are For Idiots -- And Here's The Math To Prove It" (April 9, 2009), MediaPost, http://www.mediapost.com/publications/?fa=Articles.showArticle&art_aid=103729

against each other for publisher inventory, but it is also possible for publishers to sell inventory to other publishers (who then resell it as part of a larger block), or networks to sell to other networks. Thus there is a breakdown of the simple model where advertisers buy from networks, who in turn buy from publishers; now everybody is buying from everybody else.

Yahoo runs two ad exchanges: Right Media and APT. Right Media was acquired by Yahoo in 2007, and claims to handle over 8 billion impressions per day, for premium as well as mid/long-tail advertisers and publishers. This is all auctioned in real-time -- there is no provision for reserving inventory ahead of time.   APT, on the other hand, appears to be Yahoo's vision of the next generation of exchanges. It allows both real-time and “futures” trading, also known as “non-guaranteed” and “guaranteed” inventory. “Guaranteed” inventory is bought ahead of time: the seller guarantees a fixed number of impressions or clicks during a specified time period. “Non-guaranteed” inventory, on the other hand, is auctioned in real-time based on price. APT was just rolled out in late 2008, and currently is only offered to a consortium of newspaper publishers affiliated with Yahoo. However, the company appears to be planning for APT to replace Right Media in the long term[11].

Interestingly, both Right Media and APT offer public application programming interfaces (APIs) which allow third-party software developers to develop applications that manage ad campaigns in the exchanges. These third-party systems can create ad campaigns, set bid prices, offer publisher inventory, and access campaign performance data. The APT interface even allows agencies and networks to manage multiple advertiser accounts within the system, as well as define approval workflows for ad creatives. These APIs potentially allow third parties to develop innovative services to help advertisers, publishers and networks manage their business through the exchange, and not be limited to the user interface and tools provided by Yahoo.

The exchanges reflect a standardization of some of the parameters and interfaces of advertising buying and selling. As long as networks can comply with these standard interfaces, they are free to innovate in terms of how they buy inventory and package audiences. Likewise, publishers must characterize their audiences according to standardized criteria, such as demographics (age, gender, household income), so that advertisers and networks can compare publishers to each other and decide which offer the best inventory (the means for determining these demographics will be discussed further in the next chapter). However, the exchange does not force advertisers and publishers to work with all of the networks in the market (or vice versa); each of the parties is still free to choose with whom it does business.

There is another kind of player in the ecosystem, the *yield optimizers*, which in some ways are similar to ad exchanges, but are more specifically focused on serving publishers by ensuring that the networks that can pay the publisher the most are showing ads. Like an exchange, the yield optimizer runs an auction for each impression in real-time to determine the highest price for the sale. Two prominent yield optimizers are Rubicon and AdMeld. AdMeld marketing material reports that it can increase publisher revenues from ad networks by 30 to 300%. (Note that yield optimizers do not intervene in the direct sales process, so they do not increase yields from direct ad sales.) The Huffington Post, a popular news website, reported an increase in ad network revenue by 200% from the AdMeld system. Prior to using AdMeld, they had to negotiate with

[11] Yahoo, “APT from Yahoo! FAQs”, http://apt.yahoo.com/faqs.php, accessed Aug 3, 2009

individual networks on an ongoing basis to determine which networks would provide the most revenue for particular sections of the site (and perhaps for particular users as well).

2.6 ISPs

It is worth commenting about the role of Internet service providers (ISPs) in this ecosystem, as this is potentially a shift in the landscape. ISPs have largely been absent from the online advertising ecosystem, although there have been cases where they replaced the ads on publisher sites with ads of their own choosing. In addition, there have been attempts by ISPs to manage the placement of ads on websites, most recently with abortive partnerships with the company NebuAd to do behavioral targeting based on ISPs' extensive visibility of online traffic (Ohm, 2009). ISPs soon cancelled these projects because of public criticism of them as an invasion of users' privacy; there were Congressional hearings about the matter, and a class-action lawsuit against NebuAd. However, it is worth considering whether ISPs may yet try to enter the online advertising space in some other way in the future.

In the US, most ISPs were either traditionally cable operators or phone companies, but those boundaries have broken down as both types of companies offer a "triple play" of phone, TV and Internet service. Consumers increasingly want to watch video online, which demands greater bandwidth from ISPs with challenges the traditional advertising monetization model of TV. For example, NBC, News Corp. (Fox) and now Disney (ABC) have a joint venture Hulu to put TV shows online. It is estimated that NBC gets only 10 cents in revenue from online video (from sites such as Hulu) for each dollar in broadcast (O'Leary, 2009). Partly this is because agency media buyers are still trying to understand how to buy inventory on Hulu, where the standard Nielsen TV ratings do not exist, and advertisers have limited knowledge of which shows include their ads. But Hulu may also be intentionally withholding inventory from the market to avoid cannibalizing their regular TV ad sales.

Convergence also manifests itself in other ways, as NBC has separate deals with both Google and Microsoft to use online systems to sell some of its ad inventory[12]. This comes after a failed attempt by eBay several years ago to build an auction marketplace for cable ad inventory.

Canoe Ventures is an initiative between all of the major cable operators to standardize the technology for deploying targeted and interactive advertising, so that advertisers can launch campaigns on all of the cable systems via one process, rather than having to work with different systems for different operators. It is not attempting to build a single marketplace (like an ad exchange) where ads are bought and sold, rather it is just defining the technical protocols by which advertisers, cable operators and networks can communicate. Detailed technical specifications are available online for download[13]. Initially, the hope was to deploy an initial version of a targeted advertising system this year, where a special version of an ad could be deployed to specific high-income cable zones -- but this ran into technical and business

[12] Chris Albrecht, "NBC Enters Into Targeted TV Ad Pact with Microsoft" (June 18, 2009), NewTeeVee, http://newteevee.com/2009/06/18/nbc-enters-into-targeted-tv-ad-pact-with-microsoft/, accessed August 17, 2009

[13] http://www.advancedadvertising.tv/

difficulties with existing cable infrastructure (Spangler, 2009). In the near term, Canoe is focusing on an interactive TV product which will allow viewers to click a button on their remote controls in order to receive more information about a product.

Time Warner and Comcast have also launched an initiative called "TV Anywhere", currently in trial phase, which allows cable subscribers access to TV shows online[14]. In other words, the channels a subscriber watches through cable TV, for example TNT and TBS, would also provide their shows online via a yet-to-be-revealed subscriber authentication mechanism. Content would be accessible via any broadband connection, not just the connection in the subscriber's home. Thus he or she might access the content at a cafe or via a mobile phone after authenticating. Also, the content would be accessible from a number of sites, such as comcast.net, fancast.com, TNT.tv, and TBS.com; viewers would not be forced to go to a single source to access the content.

There is another kind of player in this market, which is in some ways like the ISP, but in other ways different. The content-delivery network Akamai is now also offering a CDN-based behavioral advertising product. Akamai is not an ISP, but it is in some sense "inside the network" so it has some knowledge of people's browsing habits. They also know some of the content that people are seeing because they serve the content. On the other hand, like an ad network, the CDN can only identify users based on cookies, not their physical connection to the Internet (which will be discussed further in the technology chapter). In addition, CDNs would not be covered by the privacy law that governs ISPs, the Electronic Communications Privacy Act (ECPA).

2.7 Other players: data exchanges and web analytics services

Finally, it is worth noting a couple of other kinds of players in this system just as examples of how the value chain is evolving and differentiating.

There are now "data exchanges" named BlueKai and eXelate which buy and sell information about individual users' interests. For example, a car-shopping website which observes that a particular user is looking at information about the Toyota Prius can sell that information to these exchanges, which in turn sells the data to marketers. Thus, data exchanges represent a separation of the data about individuals from the ad targeting.

There are also *web analytics* services which collect data about user activity on websites in order to help the website designers better understand their users. Google Analytics, Omniture (recently acquired by Adobe), and Webtrends are three prominent third-party web analytics services. Publishers send these services records of their users' click patterns: which pages each user visits and in what order. The analytics services look for patterns in this data in a variety of ways, such as identifying the most common navigation paths through a website, and the places where users most frequently leave the site. But, one can imagine this data also being used for marketing purposes.

[14] Chris Albrecht, "Comcast and Time Warner Talk TV Anywhere, But Don't Say Much" (June 24, 2009), http://newteevee.com/2009/06/24/comcast-and-time-warner-talk-tv-everywhere-but-dont-say-much/, accessed August 17, 2009

2.8 Actors that play multiple roles: Google, Yahoo and AOL

Finally, it is worth noting the existence of actors that have multiple roles in this ecosystem: most notably, Google, Yahoo and AOL.

The Google search engine can be viewed as a publisher, in that it provides content to users and receives revenue from displaying ads against that content. It is the dominant search engine in the US, and the leading seller of search advertising. Google also operates two ad networks: Google AdSense and DoubleClick. Google AdSense places text ads (the same ads seen on the search engine) on the pages of many other publishers. DoubleClick is a display ad network also reaching many users and publishers. Finally, Google recently announced the DoubleClick ad exchange for display ads.

Yahoo, like Google, has a search engine (albeit with much less traffic than Google), ad network and an ad exchange. It is also a portal, a particular type of publisher with a wide variety of other kinds of content and ways to engage users. Finally, AOL has a portal and ad network, but not an exchange.

Chapter 3. Advertising measurement, behavioral targeting and ad pricing

This chapter examines behavioral advertising in more detail, and also how online ads are priced. But first it looks at general audience measurement to illustrate the kinds of data available to marketers about websites, and the importance of independent third parties that provide credible numbers about viewership and usage. In this respect, online advertising is following the path of TV and other media, with their Nielsen ratings. However, the Internet presents unique measurement problems and opportunities. It is conceivable that behavioral data analysis will become an alternative way to produce general audience measurements.

3.1 Online general audience measurements / panel surveys

General audience measurements are the baseline or summary measurements that give marketers an overview of the people who visit a given website, in order to plan their ad campaigns. These may be seen as the Internet analogue of the Nielsen ratings that are fundamental to television advertising. These figures include the number of unique visitors to a website, the length of time spent on the website, and the demographics of the visitors (age, gender and income level). As with TV ratings, it is desirable for these figures to be produced by independent third parties not associated with particular publishers, because publishers have an incentive to report large numbers of viewers in order to attract greater ad revenue (Bermejo, 2007). As will be discussed further in the technology chapter, the publisher does have some idea of the number of visitors to its site, but even its data has limitations. Also, the publisher can not learn the demographics of those visitors without asking for demographic information as part of the registration process (and relying on the visitors to provide accurate information). Thus the independent measurement companies can provide valuable information to publishers as well as advertisers.

The measurement companies generally use survey panels, large numbers of people who have agreed to install software on their computers that records all of their browsing activity and shares it with the survey company. These people also report their age, income and other demographics, so that the survey company can produce statistics about the comparative demographics of different sites. Nielsen and comScore are the biggest names in online audience measurement, but there are other players, such as Quantcast, Compete and recently Google's free Ad Planner tool. All of these companies can estimate, for example, the percentage of a given site's visitors that are males between 18 and 25, or have an annual income above $100k.

One difficulty with these kinds of survey-based measurements is that it is hard to produce data for sites with small audiences (also known as the problem of "audience fragmentation"). Suppose that it is necessary to have at least 100 people in the panel visit a site in order to produce statistics about that site. In that case, a panel size of 200,000 could only measure sites with more than 100,000 visitors[15]. ComScore claims to have more than 2 million people worldwide in its panel[16]. Nielsen has recently enlarged its US panel from 30,000 to over 225,000 people, allowing it to measure the audiences of 30,000 distinct sites[17].

Another issue with panel measurements is that it is difficult to collect data on browsing behavior in the office, because office IT departments may not allow the installation of the monitoring software. Finally, there is the challenge of projecting and weighting survey results to account for the biases in the survey panel. Given all of these issues, it is perhaps not a surprise that different companies can produce significantly different estimates for the same site's audience. For example, in March 2009, Nielsen and comScore reported dramatically different viewership numbers for the online video site Hulu: Nielsen reported that Hulu had 8.9 million viewers, whereas comScore reported 42 million[18].

On the following pages are screenshots of the audience statistics for the *New York Times* website from two different free analytics services: Google Ad Planner and Quantcast[19]. In both cases, the data refers only to viewers coming from the U.S. The trend graphs show an estimate of the unique number of visitors each day, over the past two years. The two services show similar results, although Quantcast reports slightly higher numbers, especially recently (1.9 million daily visitors for Quantcast, compared with 0.7 million for Google). Interestingly, the numbers are reversed when looking at unique visitors per *month*: Google shows 19 million, whereas Quantcast gives 14 million (not shown in the figures). In terms of the demographics, Quantcast reports that men dominate the readership, whereas Google reports a fairly even gender distribution. Quantcast also reports a slightly richer audience than Google (a higher percentage of households with income above $100,000). It is hard to say which set of statistics is more "correct", without knowing more about the two services' methodologies.

These panel surveys can also measure how many people see particular advertisements, thus providing a way for advertisers to double-check the numbers generated by their own systems. However, again, the survey data is only reliable for advertisements targeted at large audiences and on high-traffic sites. Once advertisements are targeted to small numbers of individuals based on particular aspects of their browsing history, for example, these survey panels can no longer produce credible data. In short, behavioral targeting makes third-party verification of advertising display much more difficult.

[15] This is a rough calculation, based on the fact that there are approximately 190 million Internet users in the US, and assuming that the site's audience demographics are not skewed significantly relative to the population of all Internet users. A more statistically-savvy analyst might produce a different number.

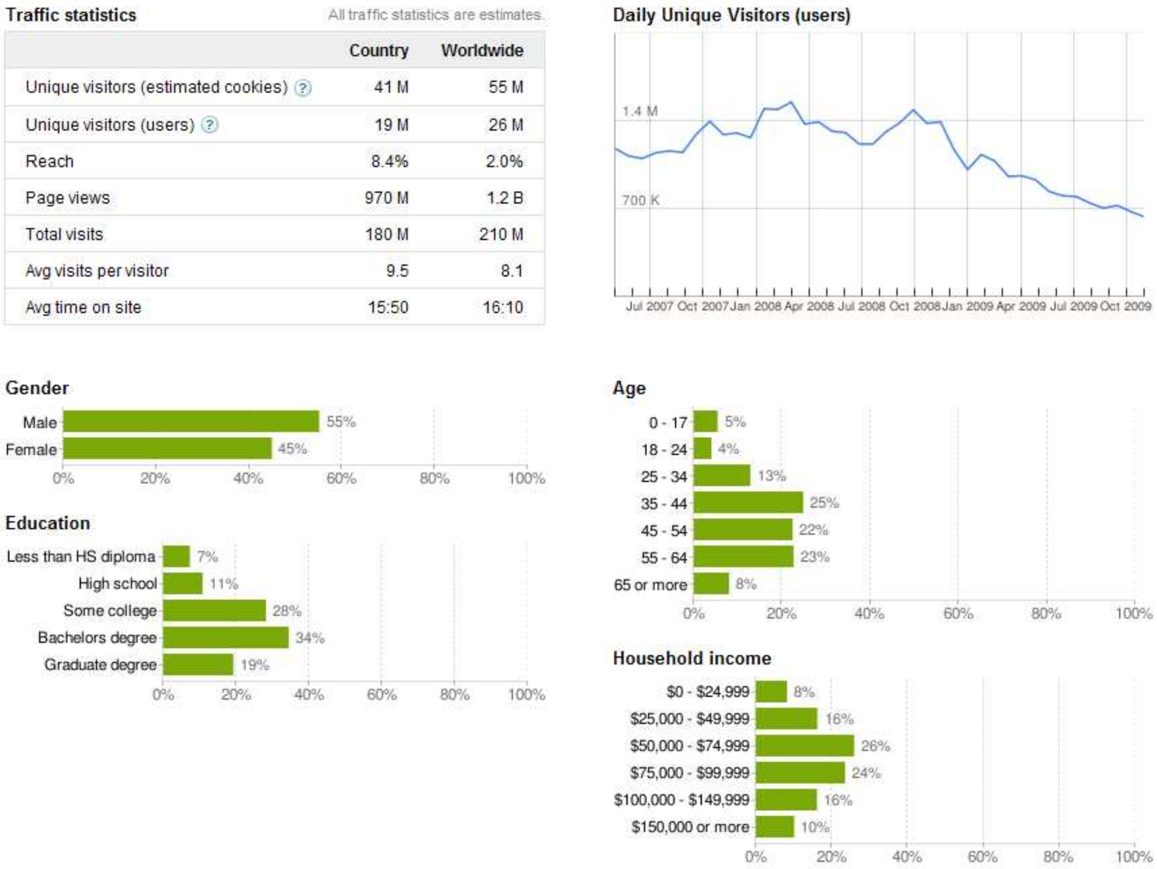
[16] comScore, "ComScore announces Media Metrix 360: the Next Generation of Global Digital Audience Measurement" (May 31, 2009), http://comscore.com/Press_Events/Press_Releases/2009/5/comScore_Announced_Media_Metrix_360

[17] Nielsen NetReporter newsletter, July 2009, http://en-us.nielsen.com/etc/medialib/nielsen_dotcom/en_us/documents/pdf/newsletters/netreporter.Par.69623.File.dat/NetReporter_0907.pdf , accessed 17 August 2009

[18] Brian Stelter, "Hulu Questions Count of Its Audience", *New York Times* (May 14, 2009)

[19] Quantcast, http://www.quantcast.com/nytimes.com and Google, https://www.google.com/adplanner/#siteSearch?identifier=nytimes.com&geo=US&trait_type=1&lp=false accessed 7 Dec 2009.

Figure 2. Screenshot of Google Ad Planner audience data for "nytimes.com"

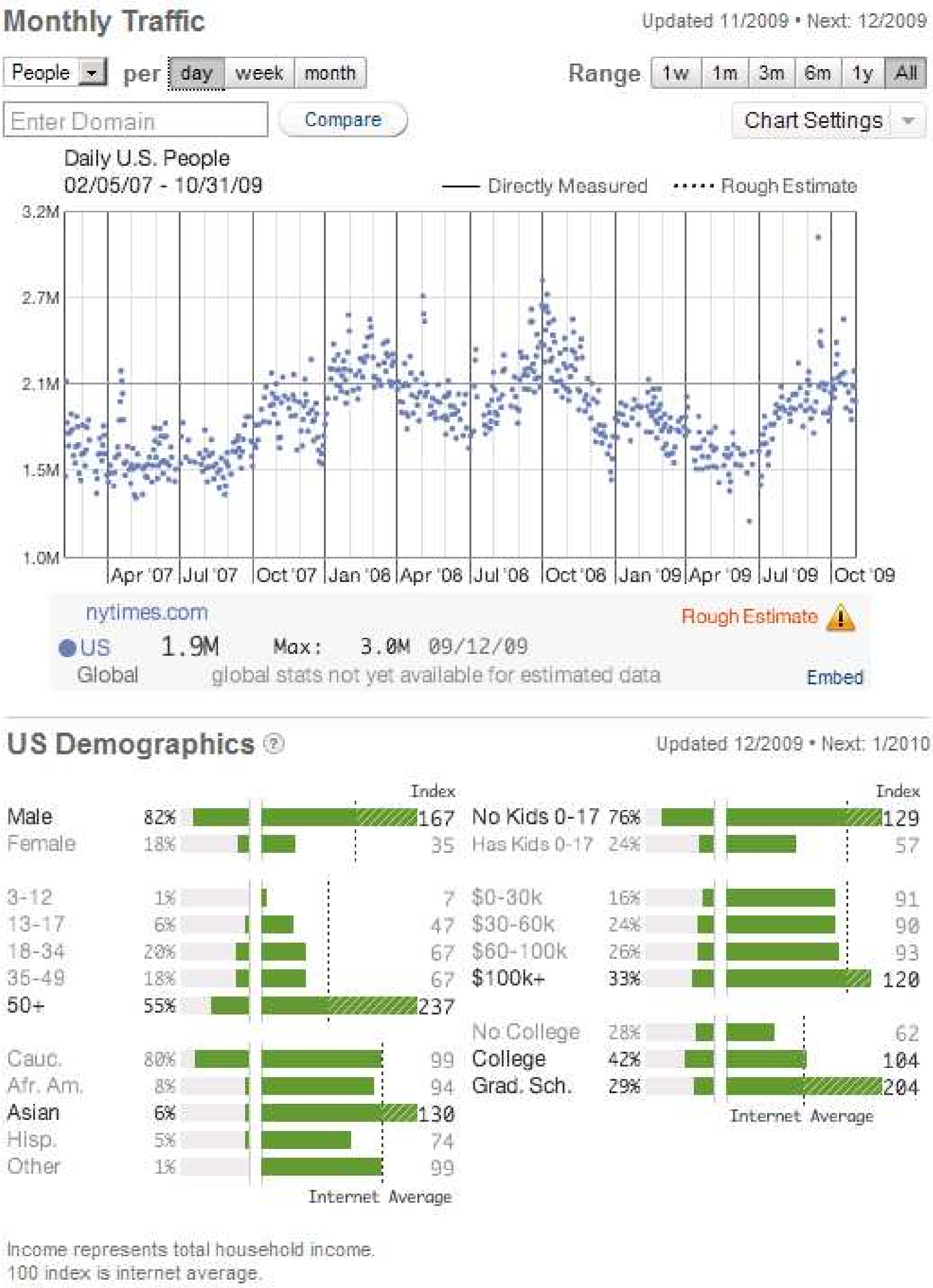


Traffic statistics

All traffic statistics are estimates.

	Country	Worldwide
Unique visitors (estimated cookies) ?	41 M	55 M
Unique visitors (users) ?	19 M	26 M
Reach	8.4%	2.0%
Page views	970 M	1.2 B
Total visits	180 M	210 M
Avg visits per visitor	9.5	8.1
Avg time on site	15:50	16:10

Figure 3. Screenshot of Quantcast audience data for "nytimes.com"



Monthly Traffic

Updated 11/2009 • Next: 12/2009

Daily U.S. People
02/05/07 - 10/31/09

nytimes.com

Rough Estimate ⚠

● US 1.9M Max: 3.0M 09/12/09

Global global stats not yet available for estimated data

US Demographics

Updated 12/2009 • Next: 1/2010

		Index
Male	82%	167
Female	18%	35
3-12	1%	7
13-17	6%	47
18-34	20%	67
35-49	18%	67
50+	55%	237
Cauc.	80%	99
Afr. Am.	8%	94
Asian	6%	130
Hisp.	5%	74
Other	1%	99

		Index
No Kids 0-17	76%	129
Has Kids 0-17	24%	57
$0-30k	16%	91
$30-60k	24%	90
$60-100k	26%	93
$100k+	33%	120
No College	28%	62
College	42%	104
Grad. Sch.	29%	204

Income represents total household income.
100 index is internet average.

3.2 Ad targeting

The problem of matching advertisements with advertising spots is called ad targeting. The advertising network and advertising exchange faces what is ultimately a decision problem: given a webpage, user, and a set of candidate advertisements, which advertisement will yield the most profit, subject to contractual guarantees and constraints (i.e. "this ad must be shown two million times in the next week"; "do not show this ad on sports websites")? The decision is generally based on some combination of a number of factors: the content on the webpage, the time of day or time of week, and whatever is known about the user, which might just be that the user is coming from California, or might be much more extensive (Broder & Josifovski, 2009). A *contextual* ad targeting system is one in which the webpage content is the primary determinant of the ad that is displayed, whereas a *behavioral* system is one in which the past behavior of the user is more important than the content of the page. In practice, systems often combine contextual and behavioral targeting, making use of whichever approach is suitable for each ad impression; if very little is known about the user, then a contextual approach may be necessary, but if the user is well-known, then a behavioral approach is possible. For example, the Google AdSense ad network initially used purely contextual targeting, but now uses a combination of contextual and behavioral data.

Contextual targeting refers to targeting of ads based on the context of the ad. The goal is to identify the sites that have the greatest "fit" with the ad in terms of their content. Webpages are analyzed to identify the topics represented on the page, and ads are selected which are relevant to those topics. Bamboo ads go on bamboo pages; fishing ads go on fishing pages. This kind of targeting is still very common.

Behavioral targeting refers to targeting focused on individual users and the interests indicated by their browsing activity. The idea is to develop profiles of users based on their activities across a number of sites. For example, the tracking system may observe that a person visited the "Toys" section of Amazon.com (an online retailer), looked at several different products, made a short detour to the "DVD" section, and then visited another website Toyforum.com where parents discuss toys and offer their opinions about which kinds of toys are better for kids. A behavioral tracking system may also record the frequency with which users visit a site or particular sections of a site; some users may visit the site just once a year when shopping for Christmas gifts, whereas others may visit the site repeatedly as they consider making purchases throughout the year. Based on all of this information, other websites may then present that person with an ad for a particular toy, even sites that do not have anything to do with toys or children. Behavioral targeting can also take the simple form of displaying an ad about a product to a person who has already expressed interest in that specific product; this is referred to as "re-targeting".

Behavioral targeting may be viewed as an extension of database marketing and direct marketing techniques that developed in the 1970s and 1980s (Turow, 2006). The increased availability of computer technologies for business led to the rise of a new kind of direct marketing, based on "segmenting" consumers into various categories in order to send customized messages to each group. Databases of income levels, age, race, spending patterns, and lifestyle preferences for ZIP codes or even individual households became available. This data could be used to segment potential or existing consumers in a variety of ways. Before the Internet, companies like Equifax and Claritas were already segmenting American households into a range of psychographic categories (such as “tree hugger” and “bible thumper”) that marketers used to craft customized messages. Companies also built databases of individuals’ purchase activity and credit history. All of this data can in fact be used to target online advertisements, once the advertising network (or whoever is placing the ad) has access to a street address, account number or other connection to an offline identity (Winterberry Group, 2009).

As noted earlier, behavioral advertising comprises two separate activities, which need not be done by the same company or at the same time: observation of users’ behavior, and targeting ads based on that behavior. As the next chapter will describe in more detail, observation or tracking can occur without the display of an ad, and likewise ads can be displayed without behavior being tracked. In addition, the company which is collecting data about users may not be the same company that is targeting ads with that data. The existence of “data exchanges”, e.g. BlueKai and eXelate, which collect behavioral data from a number of publishers and resell that data to advertisers, demonstrates how these two activities can be separated. These exchanges may be seen as the online analogues of the mailing list brokers for offline direct marketing. In an industry conference[20], some speculated that data about users could in fact have more value than ad inventory.

The attached table is a list of the categories which are used to classify users in AOL's advertising system. In marketing parlance, these categories are called “segments”. A bit of terminology is perhaps apropos: “intender” refers to a person who is estimated, based on their browsing behavior, as being close to the point of making a purchase. Thus showing an ad to an “intender” might be especially likely to influence a purchase decision.

[20] The *digiday:NETWORKS* and *digiday:TARGET* conferences in New York, June 2009. See also http://www.digidaynetworks.com/ and http://www.digidaytarget.com/.

Table 3. List of behavioral segments offered by AOL[21]

- Academic-Minded
- Active Gamer
- Affluent
- Apparel Shopper
- Auto Intender
- Auto Intender Custom-Competitive Set
- Auto Intender-Crossover
- Auto Intender-Hybrid
- Auto Intender-Luxury
- Auto Intender-Midsize
- Auto Intender-Minivan
- Auto Intender-Pickup
- Auto Intender-Sedan
- Auto Intender-Sports Car
- Auto Intender-SUV
- Auto Intender-Used
- Auto Parts Shopper
- Avid Golfer
- Black Voices Audience
- Black Voices/Auto Intender
- Black Voices/Die Hard Sports Fan
- Black Voices/Entertainment Buff
- Black Voices/Money Minder
- Black Voices/Moviegoer
- Black Voices/Television Watcher
- Black Voices/Traveler
- Black Voices/Women Audience
- Born to Budget
- Business Decision Maker
- Business IT Influencer
- Business Traveler
- Career Watcher
- Casual Diner
- Computer Intender
- Die Hard Football Fan
- Electronics Shopper
- Entertainment Buff
- Environmentally Minded
- Family Chef
- Family Planner
- Geared for Games
- Health Seeker
- Healthy Moderation
- Home Decor Shopper
- Home Improvement Shopper
- Insurance Intender
- Investors
- Latino Audience
- Latino/Auto Intender
- Latino/Die Hard Sports Fan
- Latino/Entertainment Buff
- Latino/Money Minder
- Latino/Women Audience
- Mobile/Wireless Intender
- Money Minder
- Mortgage Intender
- Motor Sports Fanatic
- Motorcycle Intender
- Moviegoer
- Moviegoer - Action/Adventure
- Moviegoer - Comedy
- Moviegoer - Family & Children
- Moviegoer - Horror
- Moviegoer - Sci-Fi
- Music Enthusiast - Country
- Music Enthusiast - Hip Hop/R&B
- Music Enthusiast - Pop
- Music Enthusiast - Rock
- News Follower
- Outdoor Sportsman
- Pet Lovers
- Primed to Purchase
- Ready for Showtime
- Real Estate Intender
- Retirement Planner
- Small Business Owner
- Style Maven
- Sweepstakes
- Technology Maven
- Traveler
- Traveler - Cruises
- Traveler - Flights
- Traveler - Hotels
- Traveler - Rental Cars
- Trendy Homemaker
- Tuned to Travel
- Wired for Electronics

The AOL segments are likely to be relatively large (on the order of hundreds of thousands or potentially millions of people). However, other systems may segment users into much smaller categories, sometimes called "microsegments" or "nanosegments"[22], which might comprise only a few thousand people.

The following section will examine in more detail the methodology underlying behavioral targeting – except for how the data is collected, which is explained in Chapter 4.

3.2.1 How behavioral targeting is implemented

Generally, behavioral targeting makes use of statistical techniques, in a couple of ways. One is clustering, which groups users into categories based on common aspects of their browsing

[21] Source: AOL, http://www.platform-a.com/advertiser-solutions/audience-targeting/behavioral-targeting/audience-behaviors, accessed July 17, 2009

[22] For an example, see DataXu, a startup company offering a "real-time bidding" platform. http://dataxu.com/benefits.php

behavior. Then if you observe that some of the users in a category have clicked on an ad, the system might display the same ad to the other users in a category. Also, the system can look for correlations between users' browsing behavior and ad clicks or purchases[23]. Any time a user sees an ad, the user's response (or lack thereof) to that ad can be recorded and added to the user's profile. The system then tries to identify what aspects of the users' browsing behavior are most strongly correlated with clicking or purchasing. These correlations could be relatively simple, such that people who spend a lot of time on health websites are most likely to purchase particular home health care products; but they could also be more complex correlations involving multiple aspects of browsing behavior. As a random example, the system might notice that people who read about health-care reform and herbal teas are more likely to click on ads for low-risk mutual funds. This would suggest some new kinds of customers for the mutual fund company to consider, or at least some new places to advertise.

One may categorize behavioral targeting systems according to whether they are *category-based* or *unstructured*. A category-based system allows advertisers to specify particular categories of people they would like to reach, by choosing from a list like the one from AOL. An unstructured system, on the other hand, simply looks for statistical correlations between browsing behavior and a particular desired outcome, such as clicking on an ad. The system attempts to "learn" what kinds of browsing behavior are associated with a click, and need not explicitly classify users into a categorization scheme like the one given above. An advertiser would not (and could not) specify what kinds of people he is interested in reaching. The system would just assume that people that click on the ad are the people the advertiser wants to reach. Also, one can imagine *hybrid* systems which first classify users into categories, and then use unstructured methods to distinguish between different users in a category. In other words, Dick and Jane might both fall in the category of "Health Seeker", but be interested in different kinds of health-related ads. A hybrid system would show them different kinds of health-related ads based on the specific kinds of health websites they visit, whereas a pure category-based system would be limited in the kinds of distinctions it can make between users. The ultimate extreme of an unstructured system is one that produces a different set of profiles for each ad. It then takes some 'training' for each individual ad to identify the kinds of behaviors associated with clicks on that particular ad. For example, each pharmaceutical ad might have a different profile of associated behaviors.

The distinction between category-based and unstructured systems will be important later, when discussing the kinds of control users might exercise over ad targeting. It also influences how the behavioral targeting system might be used. With a category-based system, lists of users falling into particular categories may be sold to any interested party, much in the way that mailing lists are sold for direct marketing purposes. Category-based systems thus enable the functional separation of data analysis and ad placement, whereas unstructured systems are more suited to an integrated business model where data analysis and ad placement are performed by the same party.

[23] There are other kinds of data that could conceivably be used for behavioral targeting, such as individuals' postings on blogs, conversations on social networks, and other kinds of user-generated content. Simply knowing who an individual is talking to online may be useful information, to understand how networks and communities influence individuals' consumption decisions. But it does not appear that targeting systems are currently making use of such information.

It is also conceivable that behavioral activity could be used to predict demographic variables such as age and income, so that ads can be targeted based on such demographics even when it is not possible to ask every user to report his or her age and income (Hu, Zeng, Li, Niu, & Chen, 2007). In other words, if a user visited many websites oriented towards seniors, the system might guess that the user falls into a 65+ demographic. This is one example of category-based targeting. It also indicates how behavioral data could be used to produce general audience measurements for websites without the sampling biases of a survey panel; all visitors to a website could be classified based on their browsing behavior.

3.3 Pricing

There are several ways of pricing online advertisements. One common method is CPM or cost-per-mille impressions, where an "impression" is one viewing of an ad. In other words, payment is based simply on the number of people that see an ad. Another common payment option is CPC or cost-per-click, where the advertiser pays only when a person clicks on an ad. A CPM rate is the price per *thousand* impressions; thus, a $3 CPM means that for every thousand people that see an ad, the advertiser pays $3. On the other hand, a CPC rate is the price for a single click. For both CPC and CPM, quoted figures tend to be of the same order of magnitude -- single-digit or at most double-digit numbers of dollars. For example, the self-service advertising system of the *New York Times* offers a starting CPM of $8, which increases in $2.50 increments as the targeting becomes more specific (e.g., a $10.50 CPM to target California residents)[24].

The choice of CPM or CPC is usually based on the advertiser's objectives for the campaign. An advertiser who is simply interested in increasing awareness of its brand may prefer to pay on a CPM basis, because its goal is to have large numbers of people see the ad. On the other hand, if the goal of an ad campaign is to get consumers to buy a product, register on a site or perform some other kind of action, the advertiser may prefer to pay on a CPC basis. A further elaboration of the CPC payment model is the CPA or cost-per-action model, in which the advertiser pays only when a person performs some action (such as registering or purchasing) on the advertiser's site.

When viewed in terms of the risk taken by the network or publisher receiving the ad payment, a CPM arrangement gives the lowest risk, because the network is guaranteed to receive revenue once ads are displayed. CPA gives the greatest risk for the network, because payment is contingent upon factors which are to some extent beyond the control of the network, such as users' interest in the product being advertised. The risk level of a CPC arrangement lies in the middle: it is riskier than CPM for a network, but not as risky as CPA. If the network has some historical data indicating which users are more likely to click on a given ad, it may be able to reduce this risk.

One report estimates that about 57% of 2008 ad spending (or about $13.3 billion) was priced on a performance basis, i.e. CPC or CPA (Interactive Advertising Bureau & PricewaterhouseCoopers, 2009). Given that most or all of the spending on search advertising and lead generation is CPC, this would suggest that most display advertising was priced CPM.

[24] New York Times, http://www.nytimes.com/marketing/selfservice/help.html, accessed June 17, 2009

ComScore estimates that U.S. Internet users viewed a total of 4.5 trillion display ads in 2008, or about 2,000 ads per month for the average user (comScore, 2009). Given that about $6.5 billion was spent on display ads, the average CPM across all display ads would be $1.44. By comparison, another report estimated average CPMs for newspapers, magazines and TV to be in the $5 to $10 range[25]. Thus, on a CPM basis, display advertising is on average less expensive than its offline equivalents. However, the averages obscure a great amount of variability. TV networks may receive CPMs in the $40 to $90 range for major sporting events[26]. Likewise, targeted online advertising can yield higher CPMs. For example, LinkedIn, a networking site for professionals, quotes CPMs in the $60-$70 range for ads targeted to corporate executives, IT professionals, and other categories of users[27]. On the other end of the spectrum, publishers working through ad networks may receive CPMs below $1[28].

There is little data on how much of the total ad spend is taken by networks, and how much is passed on to publishers, but one network states publicly that it shares 70% of gross revenues with publishers[29].

Table 4 gives an idea of revenues that several leading publishers received over the course of a year for display and video ads. Note that the ad revenue from a thousand average users over the course of an entire year is not much higher than the market-wide average CPM of $1.44, because the average user views just a few ads. Presumably there is a small subset of users that generates most of the ad revenue.

[25] eMarketer, "US Advertising CPM, by Media, 2008" (Feb 1, 2009), accessed November 22, 2009. Source: Jefferies & Company, Media Dynamics

[26] eMarketer, "Average Network TV Advertising Pricing for Major US Sporting Events, 2008 (thousands and CPM)" (Mar 10, 2009), accessed November 22, 2009. Source: TNS Media Intelligence

[27] LinkedIn advertising rate card, http://download.linkedin.com/corporate/advertising/pdf/pdf_ratecard.pdf?goback=%2Emml_inbox_none_DATE_1%2Eail , accessed 29 September 2009

[28] eMarketer, "Average Advertising Network CPMs for US Websites, by Size, February-July 2008" (Aug 19, 2008), accessed November 22, 2009. Source: PubMatic

[29] Casale Media website, "The Network Model: A Continuous Value Cycle", http://www.casalemedia.com/network_model/, accessed June 17, 2009

Table 4. Display and video advertising metrics for leading publishers, Dec. 2006 - Nov. 2007[30]

	Advertising revenues (millions)	Unique users (thousands)	Ad revenue per 1000 users	Page views (thousands)	Ad revenue per 1000 page views
Yahoo!	$1,375.90	108,734	$12.65	33,425,115	$0.04
MSN	$422.80	95,594	$4.42	14,764,863	$0.03
AOL Media Network	$286.60	91,303	$3.14	7,836,853	$0.04
MySpace	$480.10	57,784	$8.31	30,900,015	$0.02
Weather Channel	$78.80	36,844	$2.14	900,176	$0.09
About.com	$35.80	35,948	$1.00	304,741	$0.12
MSNBC	$250.80	29,230	$8.50	727,221	$0.34
CNN	$71.70	29,144	$2.46	1,204,612	$0.06
IMDb	$78.80	20,653	$3.82	700,601	$0.11
ESPN	$136.20	17,371	$7.84	901,889	$0.15

A recent analyst report argued that behaviorally-targeted ads account for a relatively small portion of the total online advertising spending, estimating it to account for just $0.78 billion of ad spending in 2008 (Hallerman, 2008). Yet, a 2007 survey of marketers revealed that 80% of respondents believed that behavioral targeting was an important marketing tactic, and in a separate survey, 75% of advertisers and agencies reported that they used behavioral targeting. Also, it is likely that the behaviorally-targeted ad spend is now much higher, because Google is now using behavioral targeting in its search advertising. The report quotes a potential $120 CPM for a behaviorally-targeted ad, compared to $10 for a non-targeted ad (presumably on a premium website, not a long-tail site) – a factor of 12 increase.

For another comparison point, mailing lists used for direct mail campaigns and other kinds of customer-relationship marketing can fetch prices in the range of $100 to $300 per thousand entries (Direct Marketing Association, 2009) -- a metric which may be compared to CPM. Of course, once the marketer buys the mailing list, he or she can use each address any number of times, without any extra cost. But if the marketer is buying online advertising on a CPM basis, he pays for each additional impression. So even if the behavioral targeting CPM is slightly lower, the total costs of the direct mail campaign and the online campaign could be comparable, if the marketer wants the online users to see an ad several times (or several different ads).

On the other hand, comparing behavioral targeted ads with non-behaviorally targeted online ads, the total cost of the behavioral targeting campaign could be cheaper, because while the CPM is increased, the number of viewers (and thus the number of impressions) could decrease by a much greater factor. If behavioral targeting reduces the target audience from 100 million to just 1 million people, the total cost would still be less even if the CPM increases by a factor of ten. For

[30] eMarketer, “US Online Display and Video Advertising Metrics for Leading* Online Publishers, December 2006-November 2007” (Feb 4, 2008), accessed November 22, 2009. Source: Nielsen Online, JPMorgan and industry estimates analyzed by OMMA Magazine cited by MediaPost, February 4, 2008

this reason, it is possible that behavioral targeting enables advertisers with smaller budgets to compete with larger advertisers. In any case, large advertisers may be more interested in reaching a large audience all at once with the same message, making behavioral targeting less appealing to them (Hallerman, 2008). Lastly, this simple math suggests that publishers would have to package a large portion of their ad inventory with behavioral targeting in order to make an appreciable difference in their revenues. (A 10x increase in CPM would not help much if it only applied to a small fraction of the total inventory.)

Chapter 4. Technology

This chapter discusses the technology underlying behavioral targeting, in order to understand the kinds of “identity” that exist in behavioral tracking systems, and the tools available to users to avoid tracking. Cookies, site registration and ISP monitoring are the key technological concepts to understand. The chapter then explains which parties are involved in displaying ads and responding to ad clicks, in order to identify what data is available to the different parties.

One caveat before starting: the description here of technological means of user resistance should not be taken in any way as an assumption that these tools are actually being used. Most users are not aware of these tools. The goal here is to illustrate what would be possible should more users become aware of the tools available to them.

Table 5 summarizes the various modes of tracking that will be discussed. The key points on which they differ are: whether tracking is separate from the display of ads, the means by which users can resist tracking and/or identification, and the kind of identity assigned to users. All of the technical terms in this table will be explained as the discussion proceeds; the table is provided here as a map or framework for the discussion. As will become clear, some mechanisms allow a user to be identified as they move between multiple sites/publishers (a shared identity), whereas others only allow a user to be tracked within a single site. As soon as the user moves to a new site, he or she appears as a new user to the tracking system.

Table 5. Tradeoffs between various mechanisms for user tracking

Scenario #	Technical mechanism	Advertising integrated with tracking / data collection?	User resistance mechanism	Shared third-party identity, or publisher first-party identity?
1	Ad network	Yes	Plug-in	Shared
2	First-party disguise of ad network	Yes	Plug-in, but potentially more difficult	Publisher
3	Web bug / tracking pixel	No	Plug-in	Shared
4	Web bug / tracking pixel with first-party disguise	No	Plug-in, but potentially more difficult	Publisher
5	Back-end data sharing	No	No	Publisher
6	ISP monitoring	Yes and no (both are possible)	Encrypting all traffic, and/or anonymizing routing (Tor)	Shared (ISP identity)

4.1 Basics of web browsing

Users use software called “web browsers”, such as Mozilla Firefox and Internet Explorer, to access web sites. When the user wishes to visit a website, he or she enters the name of the website into the browser, and the browser in turn contacts the website, sending a piece of information known as a *request* or *call*. The request includes, at a minimum, an identifying string for a particular webpage, known as a URL. A URL on the *New York Times* website, for example, may be `http://www.nytimes.com/ref/membercenter/help/privacy.html`. “Http” is a standard prefix found at the beginning of all URLs. The next portion of the URL, `www.nytimes.com`, is called the *domain name* and identifies which web server stores this webpage. The remainder of the URL identifies a particular page on the website. The request may also include other pieces of data, such as an identifier for the user (which will be discussed below).

The website responds to the request with the content of the webpage, in a format known as HTML (HyperText Markup Language), which includes the text of the webpage, but not the images and videos. Rather, the HTML includes additional URLs that identify the location of the images and videos, and the browser must download them separately. These images may come from the publisher’s web server, or from a web server operated by another organization or company, such as the web server of an ad network. The browser then makes additional connections to all of the web servers containing images on the page, and downloads those images. Thus opening a single web page may involve any number of separate connections to separate websites. The domain name that the user requests access to, and which provides the initial HTML webpage, is referred to as the *first-party* domain (“nytimes.com” in the above example). Any other domains that provide portions of the webpage content are referred to as *third-party* domains (such as the domains of advertising networks).

Some web browsers allow users to install third-party software, called *plug-ins*, that modify or augments the browser's capabilities in a variety of ways. Among other things, plug-ins can prevent many kinds of advertising from appearing on the user's screen[31]. This flexibility or *generativity* (Zittrain, 2006) in the web environment arises from the fact that the web is built upon open communications protocols that are not controlled by any single company or actor. Openness allows the development of browsers like Mozilla Firefox that in turn can be extended or modified by plug-ins. If the web was based upon a closed protocol controlled by a single company, that company could control the kinds of plug-ins or extensions available to the browser, potentially reducing the options available to users and the control they could exercise over advertising and tracking.

4.2 Identity: IP addresses and cookies

If a request contained only a URL and no other information, the website would have no way to distinguish between different users or even different computers. All users would see the exact same webpage. However, the web request does in fact include two kinds of additional information used to identify particular users: an *IP address* and *cookies*.

4.2.1 IP addresses

The IP (Internet Protocol) address is a set of numbers that identifies a computer on a network, for purposes of routing information between computers. All computers on the Internet have an IP address, and (to a first approximation) all computers should have a unique IP address used by no other computer.

In practice, there are ways to masquerade and manipulate IP addresses so that many different computers "appear" to the outside world to have the same IP address. This can be done for security reasons, or to conserve IP addresses, which are finite in number and thus becoming a scarce resource. In addition, even if an IP address identifies a single computer, it will not distinguish between different users on that computer. If a family shares a computer, the mother, father and all of the children will all be using the same IP address. Thus a system that uses IP address as an identifier will not be able to distinguish between these different individuals.

Finally, IP addresses are not guaranteed to be stable long-term identifiers; they may change from day to day or even more frequently. Thus they can not be used to identify households over a long period of time.

4.2.2 Cookies

Cookies are critical to the tracking of users within and across sites. This is because without cookies, a web server has no way to distinguish different people from each other, except with their IP address, and as noted above, the IP address is an unreliable identifier. The cookie is what

[31] See the *AdBlock Plus* plug-in, at http://adblockplus.org/en/ .

gives a web site “memory” of a particular user[32].

Suppose that Alice goes to a news website and reads an article about South Africa. When Alice connects to the website, the website tells Alice’s browser to store a piece of information that looks something like “userID: 98fsglk32”. “98fsglk32” is a random string that has no meaning, except that it distinguishes one person from another. When Jane accesses the same website to read an article about polar bears, the website tells Jane’s browser to store a different piece of information: “userID: 74ahjn09”. The website can then build a database of the topics about which different users have read; in simplified form, it may look like this:

User ID	**Article subject**
98fsglk32	South Africa
74ahjn09	polar bears

When Alice next goes to the website, her browser sends the user ID of “98fsglk32” to the website, so that the site can search for this ID in its database and learn that this person already looked at an article about South Africa. It can use this information in a variety of ways, for example by showing a listing entitled “Topics You Have Recently Read About” on the right-hand side of the webpage. In Alice’s case, this list would include “South Africa”; in Jane’s case it would include “polar bears”. Note that the website does *not* know Alice or Jane’s name or any other aspect of their identities; it only knows the topics they have previously read about.

These pieces of information like “userID: 98fsglk32” are called *cookies*. Each cookie has a name (for example, “userID”) and some content (e.g. “98fsglk32”). Furthermore, cookies are associated with the domain name of a website (e.g. “nytimes.com”) such that only that website has access to the information from that cookie. When Alice goes to another site, for example Facebook, her browser does not send the user ID “98fsglk32” to Facebook. Facebook must create its own cookie for Alice, for example “userID: yucca12”, that is not shared with the news website (unless particular programming techniques are used, which will be discussed below). In addition, the process of depositing and transmitting cookies occurs entirely automatically as the user accesses web content, and normally there is no visual indication to the user that the website has assigned him or her an identity that allows for tracking. (There are ways for users to learn how they have been labeled, which will be discussed below.)

Cookies are not used only by websites such as the *New York Times* or Facebook, but also by the advertising networks showing ads on those sites. Each ad network has a separate set of cookies for the same person, so the Microsoft ad network might identify Alice as “userID: mnm156” while the Yahoo network identifiers her as “userID: ufskjb6”. Furthermore, a user can have multiple cookies from a given website or ad network. The news site may have a cookie “userID: 934280clkjs”, and another cookie “websiteColor: red”, as a way of remembering the user’s preferred color for the website. As this illustrates, not all cookies are used for distinguishing between individual users; some cookies assist in the customization of an individual user’s website experience.

[32] Actually, it is technologically feasible to design a website that has knowledge of user identity, but does not use cookies (by including a user ID as part of the URL of every page). However, very few websites are designed this way. Using cookies to track identity is much easier.

Sites that use cookies to identify users may do so in one of two ways: either by automatically generating an identifier for each user (as in the first example with the news website) or by asking users to register with the site and thereby create their own user ID. In the latter case, the cookie may contain the user's own chosen ID. Cookies are necessary in order for users to register and sign in to sites; if the browser did not use cookies, sign-in would be impossible.

4.3 Cookie management and other advanced techniques

This section discussed the various ways that users can control how cookies are used for tracking, and also some techniques publishers and ad networks can use to get around some of the restrictions initially associated with cookies.

4.3.1 User control of cookies

A key implication of the way cookies are designed is that the browser is not actually *required* to record the cookies provided by websites. The browser can in fact ignore the website's request to store the information. In the example of the news website, this would prevent the website from remembering the previously-read articles. Browsers generally give users some control over the storage of cookies through configuration options in the browser's Preferences section. By default, browsers generally accept most cookies, but if the user wishes, he or she may tell the browser to refuse all cookies, thus preventing websites from tracking them or having any "memory". However, disabling cookies altogether will generally limit the functionality of many websites, because as mentioned above, disabling cookies will prevent a user from being able to log in to most websites. For example, an online retailer would not be able to remember the items you previously looked at, or the items in your shopping cart.

In the middle of the spectrum, between full acceptance and full refusal of cookies, one may tell the browser to accept cookies only from certain websites, or to accept cookies from all websites except for specific "blacklisted" sites. However, this presents an extra burden for the user, because for every site he visits and every ad network used by those sites, he must make a judgment about whether he wants to allow cookies or not. An alternative is to use a browser plug-in that manages this task automatically for the user. In this case, the plug-in may examine each cookie that a website or ad network asks to leave in the browser, and compare it with a database of the domain names associated with "good" and "bad" cookies[33]. For example, the database may indicate that cookies from "adnetwork.com" and "usertracker.com" should be blocked, and cookies from all other domains may be allowed.

Privacy-enabling browser plug-ins are generally designed by volunteers as open-source projects, in an adversarial relationship with the publishers and networks. The plug-in developers must "reverse engineer" some of the mechanisms of how the networks track users, and must continually update the plug-ins as networks and publishers change their code and redesign their systems. Thus there is an ongoing "arms race" between the data collectors and the pro-privacy developers.

[33] For an example, see the *Targeted advertising cookie opt-out (TACO)* plug-in, http://taco.dubfire.net/ .

Even without a plug-in, browsers generally allow users to look at all of the cookies deposited by websites, and delete individual ones. Thus users may also allow websites and ad networks to leave cookies, but periodically delete them, so that websites can track them over the short term but not the long term. One study by comScore, focusing on a Yahoo cookie and a DoubleClick ad network cookie, estimated that about 30% of users deleted one of those cookies during a one-month time period (Abraham, Meierhoefer, & Lipsman, 2007).

Although websites and networks can not force users to accept cookies, they can detect that their cookies are being blocked; thus, publishers and networks do have some information about how often it happens.

One way for advertising networks to work around the problem of cookie deletion is to use a different kind of cookie, associated with the Adobe Flash program. Flash is a browser plug-in widely used for showing video or making other forms of interactive webpages. In addition, websites (and ad networks) can use Flash to store information on user's computers and then send it back to the website when the user next connects to the site, essentially duplicating the browser's cookie mechanism. These "Flash cookies" are stored in a separate location from regular browser cookies, and fewer tools are available to control them (Soltani, Canty, Mayo, Thomas, & Hoofnagle, 2009). Also, fewer users are aware of their existence.

4.3.2 JavaScript, and identity linkage between publishers and ad networks

In the above discussion of cookies, it was stated that the publisher can not see the cookies for the ad network, and the ad network does not see the cookies created by the publisher. This technical restriction would make it difficult for the publisher and ad network to share data about the user, because they would each have a separate ID for the user. Thus the publisher and network would be unable to combine their knowledge about the user's behavior into a larger composite picture. However, there is a technological way to get around this restriction: JavaScript.

JavaScript is a programming language that operates within the web browser. The publisher can include JavaScript code within the HTML page that is executed by the browser upon downloading the page. Among other things, the code may combine information (such as cookies) from multiple websites. The browser may send a request to network A, asking "what is the user's ID in network A?" Network A then returns an identifying string like "49295". The JavaScript may then include this user ID in a call to network B, effectively saying to network B: "this user has ID '49295' in network A." Network B of course knows its own ID for the user, e.g. "aj411", so now it can cross-reference users between the two networks. It now knows that the person with ID "49295" in network A is the same person identified as "aj411" in its own system. The ability to cross-reference users between different networks is critical to the functioning of an ad exchange, because it allows the exchange to tell the networks: "this user is about to see an ad; for network A, his ID is …, for network B, his ID is …; what price are you willing to pay to show the ad?" The networks can then bid against each other for the right to show an ad to the user.

However, if users install a browser plug-in that blocks the ad network cookies, the cross-referencing of the user would still be prevented, because the user would not have an ID cookie for either network. In such a scenario, it is conceivable that a network could make use of an ID associated with a publisher, one that was not blocked. In other words, if the user was signed into the Fox News website with an ID of “sopchak”, other sites could use JavaScript to access that ID and then send it to ad networks for ad targeting, so that the ad network need not leave a cookie. However, this would require a business relationship between the publisher and the network (which is certainly imaginable if the network is a vertical network associated with a publisher). It is not known how often this happens.

4.3.3 DNS aliasing: obscuring third-party servers

One way that a publisher could use a third-party ad network for displaying advertisements, while still working around plug-ins that attempt to block such ads or tracking, is to use a *DNS alias* to “disguise” the third-party ad server as a first-party server (Krishnamurthy & Wills, 2009). This involves giving the third-party server “adnetwork.com” another name like “otherserver.nytimes.com”. The request for the ad banner would then go to “otherserver.nytimes.com”, which the plug-in would be less likely to find on its blacklist; in other words, the plug-in would think that the ad banner was actually part of the content of the page. However, this would force the ad network to make use of the cookies from the publisher’s domain rather than its own cookies, and again raise the question of how to track a user as it moves from one site to another.

4.4 Tracking is separate from advertising: web bugs and tracking pixels

Thus far it has been implicitly assumed that tracking or observation of a user’s behavior occurs when an ad is displayed. But in fact, tracking and ad display are independent activities. It is possible for a behavioral tracking system to observe a user’s behavior without showing an ad, by means of a *web bug* or *tracking pixel*. These are two different names for the same mechanism, in which the HTML of the publisher’s web page includes a call to download an invisible image from the tracking system’s server (or, JavaScript code could call the tracking server, but the end result is the same). The image is a one pixel-by-one pixel image of the same color as the background of the page, so it does not change how the page appears. But the call to the tracking system’s server notifies the tracker that a user has visited a particular page on the publisher’s site, so that the tracker can record this step in the user’s browsing trail. In terms of the communications between the browser and the tracking system, the web bug is no different from an ad; the only difference is that there is no visual indicator to the user that the communication is occurring. The call to the tracking system includes whatever identifying cookies the browser has for that tracker, so the tracker can distinguish between different users.

4.5 Publisher tracking with server logs, and back-end data sharing

Websites generally keep a record, called a *server log,* of all of the activity on their sites. The log includes the URL of each page that was accessed, when it was accessed, and the IP address and cookies of the user accessing each page. If the user disables cookies for the website, the only remaining user identifier is the IP address, which as discussed above, is somewhat unreliable. However, if the site has a sign-in process to enable user customization, the user may not want to disable cookies. Thus as long as the user wishes to access personalized features of a website, he must implicitly consent to at least allowing the publisher to track and record his behavior on the site.

In addition, the publisher may share this behavioral data with a third party "behind the scenes" simply by sending the data from its servers to the third party's servers. The user's browser could not block the data transmission because it does not involve the browser at all. In this case, the user would be identified by his user ID on the publisher site. If the user moved to a different site, he would now have a different ID and there would be no way for the third party to know that these two different user IDs actually identified the same person.

Thus, third parties could aggregate behavioral data from a number of different publishers in a way that browser plug-ins could not block, but they would face the challenge of reconciling the user IDs from the various publishers.

4.6 ISP monitoring

The ISP can see all of its users' connections, every site they are connecting to, and the information they send to those sites and the information they get back – with the exception of encrypted communications, which are generally used when making purchases online, accessing bank accounts or other sensitive information. (Of course, this only applies to the ISP's subscribers; Time Warner can not observing the complete browsing behavior of a Comcast customer.) By contrast, the data available to ad networks, exchanges, publishers and agencies are all limited to a particular set of webpages (as will be discussed later in this chapter). Furthermore, unlike ad networks and some publishers, the ISP also knows the user's name and street address, making it possible to cross-reference data the ISP collects about a user with data from other sources.

The limit at one point had been on the ability of ISPs to store the massive amounts of data that was flowing through their networks. However, ISPs are increasingly considering the use of "deep packet inspection" (DPI) systems which filter and analyze this data. The current limits on ISP behavior are legal and political, as discussed earlier. Without these restraints, it is possible for ISPs to produce behavioral profiles matched to street address, which then can be linked to other databases of purchase activity by street address.

There are ways for users to escape the ISP's eyes, by routing all of their traffic through an encrypted proxy such as Tor[34], but these again are cumbersome and not widely-known.

4.7 Summary: the cookie arms race, and some empirical data

To summarize, the original design of cookies was intended to prevent multiple sites from being able to share information and cross-reference users. However, technological developments have eroded that restriction, and it is relatively easy for publishers and ad networks to cross-reference

[34] http://www.torproject.org/

users. There are tools (plug-ins) available to users to limit this cross-referencing process, and prevent their activity from being tracked by particular sites or networks. However, if a site requires sign-in to access content or services, the user has no way to limit the site's ability to track behavior. The user also has no means to prevent the publisher from sharing user behavioral data with other websites or ad networks. In addition, third-party trackers can disguise themselves from plug-ins with DNS aliasing, although this makes it more difficult to track users as they move between different publishers.

Table 5 summarizes the tradeoffs associated with the various scenarios for user tracking.

Scenario #1 is the case in which an ad network uses a third-party cookie associated with its own domain to track users. In this case, the collecting of data about the user occurs at the same time as the ad is displayed. Browser plug-ins can block the ad display and collection of data. The ad network assigns its own ID to the user, which persists as the user moves between multiple sites.

Scenario #2 is the case in which the ad network server is disguised with a DNS alias such that it appears to be a first-party server in the domain of the publisher. Again, the data collection occurs at the same time as ad display. However, the user is now identified by an ID that is publisher-specific; as the user moves between sites, his ID changes.

Scenario #3 is the case of a web bug or tracking pixel which records data about the user without displaying an ad. In all other respects, it is the same as scenario #1. Scenario #4 is the case of a web bug disguised with a DNS alias. Data is collected without an ad being displayed, but in all other respects, it is the same as scenario #2.

Scenario #5 refers to data sharing between a publisher server and a third party server without any browser involvement. Data collection is again separate from ad display, and there is no way for the user to block the data collection. The publisher user ID identifies the user.

Finally, scenario #6 refers to the case of ISP monitoring. The ISP may collect user data whether or not it displays the ads. The user must encrypt his traffic in order to avoid monitoring. The ISP has an account number and street address from which the user is connecting, a relatively strong kind of identity that persists as the user moves between different sites.

There is some empirical data on the use of third-party tracking and advertising servers by popular websites (Krishnamurthy & Wills, 2009). This study's authors have collected data on the usage of third-party services by the 1200 most popular websites[35] over the past several years, through September 2008. They then identified the top third-party service providers, measured by the number of popular websites which make use of them (see Table 6). These third-party services include not just advertising networks, but also web analytics services; in fact, among the top third-party services were the Google, Omniture and Quantcast analytics services. As of September 2008, Google services reached almost 60% of the top websites. Some of these sites use only Google Analytics, others use only DoubleClick, and others use both. No player had nearly that level of market share a few years ago; before the Google / DoubleClick merger, they were each reaching around 20% of sites.

[35] As measured by Alexa, a free web measurement service.

Table 6. Most common third-party services used by publisher sites in Sept. 2008, from Krishnamurthy and Wills (2009)[36]

Third-party service provider	**Percentage of publisher sites using third-party service**
Google (including DoubleClick and Google Analytics)	60%
Omniture	28%
Microsoft	22%
Yahoo	15%
AOL	14%
Quantcast	13%
Audience Science / Revenue Science	9%

Note that the actual penetration of third-party services could potentially be higher than these numbers indicate, because the authors may only have visited a few pages on each site. If more pages were visited, at different times and from different locations, more third-party services might have been observed. This fact may explain why these numbers are inconsistent with the comScore reach figures listed in Table 2. ComScore reports that there are around 20 distinct ad networks that reach at least half of the US internet-browsing population. In other words, all 20 of those networks serve at least one ad to half of the population in the course of a month. However, the above study would indicate that most of those 20 networks do not reach a major portion of the top websites. It is possible that the networks may only be used on particular pages that the study did not visit, or that other methodological issues prevented detection of those networks[37]. Or, perhaps some of those networks are focused almost exclusively on lesser-known sites not tested in the study. This illustrates some of the difficulties with empirical data collection for online advertising: because ad targeting is increasingly based on the complex interaction of a wide range of variables, it is necessary to test a wide range of scenarios to collect accurate data.

4.8 Ad networks, exchanges and data visibility

Technical architectures determine not only the modes of user resistance, but also the kinds of data available to various stakeholders. This section illustrates this point by means of several scenarios.

The next diagram illustrates three scenarios for the sequence of connections that are made in order to display ads and click on ads. The solid lines represent communications that occur when the ad is displayed, and the dashed lines represent communications that occur only if the ad is clicked.

[36] Data comes from the text of section 4.3 and figure 7 in the cited paper.

[37] It is possible that the study failed to detect those other networks because it is necessary to follow a series of HTTP redirects to identify the true source of the ads, and the study did not follow those redirects.

The first scenario illustrates what happens if a publisher is serving an ad that has been sold directly to an advertiser (without the involvement of a network). In this case, the publisher's web server provides the ad image to the browser. If the user clicks on the ad, the browser contacts the publisher's web server again, and this server gives the browser the location of the advertiser's website. The browser then downloads and displays a webpage from the advertiser's site.

The second scenario illustrates what happens when the ad is placed through a network. In this case, the publisher's webpage instructs the browser to download the ad image from the ad network's server. If the user clicks on the ad, the browser contacts the network's web server again, in order to learn the location of the advertiser's website. This scenario is very similar to the previous scenario, except that the network's web server provides the ad image, and not the publisher's server. In addition, note that the publisher does not know which users see which ads, because it does not make that decision.

The third scenario, involving an ad exchange, is slightly more complicated. In this case, the browser requests the ad image from the exchange, but the exchange must first contact several ad networks in order to determine which network will place the ad. Once this decision is made, the exchange then provides the ad image to the browser. If the user clicks on the ad, the browser contacts the exchange again, which in turn directs the browser to contact the ad network, which finally directs the browser to the advertiser's site. As in the previous scenario, the publisher does not know which ad the user sees.

Also note that in all of these scenarios, the advertiser does *not* know anything about the users that have seen the ad but not clicked on it, unless the publisher, network or exchange provides that data to the advertiser. Thus the technical architecture of the system determines who controls access to various kinds of data. There are scenarios where particular parties may withhold some data from other parties. But it is also entirely possible for a network to share more data with an agency or publisher than the data described above, if it feels that it is in its interest to do so.

Figure 4. Illustration of communications involved in displaying and clicking on an ad

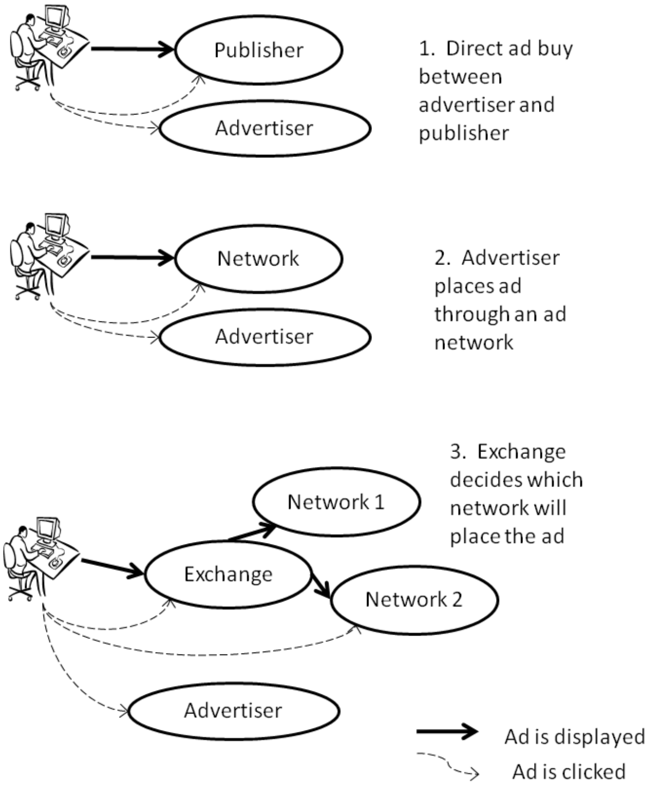

Publisher
Advertiser
1. Direct ad buy between advertiser and publisher
Network
Advertiser
2. Advertiser places ad through an ad network
3. Exchange decides which network will place the ad
Network 1
Exchange
Network 2
Advertiser
Ad is displayed
Ad is clicked

Table 7. The extent of data available to various actors

Actor / scenario	Pages and users for which data is available
Agencies—buying directly	only those who see their ads
Agencies—through networks	whatever data is shared by the network
Networks—working directly with publishers	only those who see their ads
Networks—working through an exchange	any impression the network bids on; if the network loses the bid, it knows about the impression, but not whether the user clicked
Publishers	where they place ads directly, and where networks share data
Invisible observer/tracker, using a web bug	publisher's choice of pages and users
ISPs	for their subscribers, all pages, except when encryption is used

The accompanying table summarizes the data that is available to each of the actors in this ecosystem. An agency or network might only have data about the users that saw their ads and the pages where the ad was shown. If an agency's ad is not shown on a particular page, the agency may not know which users visited that page, even if the agency placed ads directly with the publisher. The agency's only connection with the user is by showing an ad. Furthermore, if an agency is working through a network, it may not even have the complete data about everybody who sees an ad. It is up to the network to decide which data it will and will not share.

If multiple networks are competing to show ads for a publisher via an ad exchange, they may all be informed about each impression, so that they can provide a bid price for that impression. Thus all of the networks would know that a user has visited a particular page.

Publishers observe the complete browsing trails of users on their sites, but of course know nothing about what those users do on other sites. They also have complete data about ads that they place directly, but may have limited data about ads that are placed through networks.

The above discussion is entirely focused on data collection that occurs as part of the display of an ad. When a web bug is used, the data that is collected is entirely up to the choice of the publisher, who may choose to place the web bug on as many or as few pages as it likes, and for whichever users it chooses.

ISPs can observe the complete browsing history of their subscribers across all websites, except when encryption is used (for example, for sensitive transactions like online purchases, accessing medical history or financial information).

Finally, it is worth following up briefly on an issue raised in the previous chapter, regarding the difficulties with estimating the number of unique users visiting a website (Bermejo, 2007). Panel surveys are used to make these estimates, but as discussed earlier, those estimates are subject to sampling biases. It is also possible for publishers to estimate this number themselves from the cookie data recorded in server logs. However, if users delete their cookies, they will be counted multiple times, leading to overcounting. Likewise, if the same person accesses the site from multiple different computers (at home and at work, for example), he will be counted more than once. On the other hand, if multiple people using the same account on the same computer, they will only be counted as one person. Thus publisher counts of viewership have their own problems, aside from the problem that they are hard to verify independently. Estimates from third parties like ad networks and analytics services are also based on cookies, and thus have all of the same issues. It is technically possible for ISPs to make these estimates, which will be discussed further in the next chapter.

Chapter 5. Analysis and Conclusions

5.1 Introduction

In the first chapter, the following hypothesis was proposed: the demand for increasingly detailed behavioral tracking of users benefits aggregators relative to publishers because the aggregators have the economy of scale needed to provide the user data of interest. This trend would concentrate power in the hands of a small number of actors who excel at data mining. The current chapter returns to the hypothesis to assess the likelihood that the evolution of the online advertising ecosystem will result in the concentration of data in a small number of hands. In order to understand the conditions under which this hypothesis might prove correct, it is necessary to consider in more detail the relationships of the stakeholders - ad agencies, publishers, ad networks, Internet providers, and finally users – to data and the strategies these actors employ for advancing their interests. The goal here is not to prove the hypothesis true or false definitively but rather to highlight some dynamics which policymakers might monitor as indications of problems.

One general issue cuts across all of the stakeholders: the struggle for ownership of data. Data is power in the online ecosystem. Knowledge about who sees ads and who clicks on ads is valuable for marketing. The question of ownership arises wherever intermediaries, such as ad networks, place or manage ads on behalf of another party[38] (Edelman, 2009; Winterberry Group, 2009). The intermediary then has the complete picture of who sees the ads and whether they click on them, and as discussed in the previous chapter, can choose to share or not share the detailed data with advertisers and publishers (and perhaps even users). Intermediaries of any sort (ad networks or even ad agencies) who work with data on behalf of clients (both advertisers and publishers) face conflicting incentives: on the one hand, clients want to know as much as possible about where ads are displayed and who sees them; on the other hand, revealing such information could threaten intermediaries' business models – making it possible for clients to look for the same services elsewhere – and reveal proprietary information about how the intermediaries target ads.

One simple example of data not being shared is ad networks that are not transparent about where they show ads. The advertiser may tell the network to place ads on a certain category of site, but does not know which exact sites are included. Likewise, a publisher who shows ads through a network may not have control over which ads appear on the site. One assumes that advertisers want to see as much data as possible to make informed choices about ad buys and placement. Apparently, at the moment, the value provided by the networks – in terms of precision of targeting, easy access to a large audience, or easy access to a valued audience -- is great enough

[38] Some actors are addressing this issue explicitly in their marketing material. For example, see 24/7 RealMedia, "Online Publishers: Who Owns Your Audience?", http://www.247realmedia.com/EN-US/intel/research-opinions.html accessed 6 October 2009.

to compensate for the loss in transparency. But agencies are now building (or acquiring) their own networks in order to access the detailed data, among other reasons (McClellan, 2009). Furthermore, the networks have knowledge not just about who is seeing what ads, but also how much advertisers are willing to pay for them (and, the minimum prices publishers are asking for their inventory). Thus there is another information asymmetry, in knowledge about the advertising market itself.

This is not to say that there is a complete hiding of data. For example, Google's AdSense network does tell publishers which pages receive the most ad revenue. This potentially gives publishers some idea of how to design a site to gain more revenue, which benefits both the publisher and Google. As a simple example, if a publisher gains most of his revenues from the sports section of the site, he may try to promote the sports section in other areas. However, it is more difficult to imagine Google (or any other behavioral targeting network) sharing data about which users generate the most revenue, both because of privacy concerns and because it might reveal valuable information about how the network does its targeting.

The next sections look at each individual stakeholder, and explore the strategies they might pursue to improve their positions.

5.2 Advertisers and ad agencies

As mentioned earlier, there are four large ad agency holding companies which are responsible for the majority of advertising spending in the US. Thus it is appropriate to ask if these holding companies can become large aggregators with market power in their own right, given that they are starting to build their own ad networks (McClellan, 2009; Winterberry Group, 2009).

One distinction between agencies and other actors in the ecosystem is that agencies also plan campaigns in other media, such as TV, print and radio. Thus they have more knowledge about consumers' exposure to messaging across all of these different media. They are working on projects which integrate data about consumers' exposure to media across a wide range of so-called "touchpoints", including the media channels described above as well as billboards and in-store displays. Based on this knowledge they can estimate how much impact a TV ad or billboard has relative to an online ad, for different types of consumers. In this respect they are aggregating data across another dimension: that of media. It is not clear how individual users are identified in these other media forms, where no user registration or sign-in is required in order to access content; perhaps users' viewing of those media are imputed based upon geographical location, demographic or other variables.

Within the online environment, agencies can accumulate data from a variety of ad campaigns managed via a variety of networks, as well as campaigns managed directly with individual publishers. When they go through networks, they may not have access to complete data about who saw the ads and who clicked on them, only summary reports. But when the agency buys inventory directly from publishers, it presumably has access to more complete data. One question, however, is whether separate agencies within the same holding company share this data; like any large organization, there are complicated internal politics and rivalries. In addition, it is unclear to what extent the agency's clients, the advertisers, allow the agencies to take data from their campaigns and use it in combination with data from other advertisers. Advertisers benefit from analysis of data from a variety of sources, but they may also feel reluctant to share this data if it is the product of extensive research and media campaigns.

In addition, even assuming that clients and organizational politics allow for data sharing, there remains the technical question of how agencies identify who is the same user across all of these different ad networks as well as the other media channels. The same person will have one ID in one network and another ID in another network. If the agency serves ads from its own ad server, this gives yet another ID, and a publisher ad server may assign one more ID. This suggests that if consolidating data from a variety of campaigns and networks is important, there will have to be a way to connect all these disparate identities. Perhaps there are ways to connect these IDs to more persistent identities, such as site registrations, retail loyalty cards, membership cards or mailing addresses, allowing for the cross-referencing of IDs from different systems.

In short, there are ways in which advertising agencies (or, to be more precise, agency holding companies) could themselves become data aggregators with some market power, but there are also some forces which might limit the extent of that power.

5.3 Publishers

It is possible that groups of publishers within a particular topical area could band together to form their own advertising networks, in order to oppose the rise of the aggregators. In this way they could control the kinds of data which is available to advertisers about their audiences. This has already happened to some extent, with the rise of vertical ad networks focused on specific topical areas such as politics and cooking, as well as the ad networks sponsored by large publishers such as Fox. These vertical ad networks would indicate that there are economies of scale from aggregating a large number of sites within a given vertical topical area. The question is whether these economies of scale arise specifically from the ability to do behavioral tracking across a variety of sites, or from other sources; for example, vertical networks also make it possible for an advertiser to reach a large audience without have to make individual deals with many sites.

Another option for publishers is to increase their level of engagement with their users, leveraging a key advantage that publishers have. As noted earlier, users have a stronger kind of identity with publishers than with networks, because there is no way for users to give up a publisher identity without losing personalized functionality. Ad network cookies and other third-party cookies can be blocked with browser plug-ins, but blocking publisher cookies requires giving up many of the features users enjoy on websites. A site that enriches its knowledge of the user’s identity can gain more value from selling the data. This suggests that a key currency for Internet publishers will be their “depth of identity”, as illustrated by the following example, based on a hypothetical reader of the *New York Times* (although it would also apply to any other newspaper).

Suppose that the reader is required to register (for free) before accessing the site, in order for the site to develop some nominal concept of the user’s identity. The user could put anyone's name in the registration form. However, suppose he were to subscribe to the print version of the paper. Then he must give them his mailing address, and presumably he would not give them a fake address. Thus the print subscribers are a kind of “premium” customer for the *Times*, because it knows more about them. It can sell their names and addresses to direct marketers. And, if it connects his identity as a print subscriber with his identity as a web surfer, it can sell even more information. It can say that George Gupschutz on 392 Maiden Lane in Marlborough, MA is an avid Chicago Cubs fan, because he always reads the articles about the Cubs. So now, maybe the people selling Chicago Cubs memorabilia will come knocking at his door (or his computer screen). This illustrates how a connection to offline identity (i.e. the mailing address) can immediately create value for a publisher.

His browsing trail from the *Times* may also be combined with browsing trails from other websites, if the two parties agree. The *Times* could partner with a sports site like ESPN, sharing data with one another, in which case they might find that George Gupschutz reads the Cubs articles on the *Times* site, but reads about hockey on the ESPN site. Perhaps George likes the *Times'* baseball writers more than their hockey writers, and the converse for ESPN. This might give additional information and enrich the profile sold to direct marketers, about the consumption of information, and about what writers are more attractive to our consumer.

LinkedIn, a professional networking site, takes the concept of “depth of identity” even further. On LinkedIn, each user builds an individual profile focused on their career, employment history and professional interests: where they worked, for how long, what their job titles were, and their skill areas. Users have an interest in sharing this information, because hiring managers and recruiters are searching LinkedIn for people with the skills and backgrounds they need. Or, they may use LinkedIn as a kind of online rolodex for the networks they have already built, because not only does LinkedIn have their resumes, it also knows their contacts: who they have worked with, and potentially their friends as well. So employers can learn that George Gupschutz is a friend of Madeleine Zirkowski, and ask Madeleine for her opinion of George.

That explains why George would be motivated to share his information on LinkedIn, information that could have several kinds of value for LinkedIn. It could use this information to display targeted advertising; for example, a software company could pay to display an ad to only those people who are software engineers or engineering managers. It can also sell lists of users to advertisers. It may not know George’s mailing address, but it could sell those advertisers his cookie ID so that on other sites that George visits, George could see ads connected to his LinkedIn profile. In other words, he could be on the *Times* site, and see an ad that is somehow related to some piece of information from his profile (and in fact, this is already happening[39]).

Thus to increase its advertising revenue, the *Times* might consider how to increase its depth of identity. It might have to transform into a kind of “information experience”, a much more personalized news experience, which allows users to indicate which kinds of topics, which writers, and what parts of the world interest them the most. The website would then be individually customized based on the expressed interests of the user. George would see articles about Africa on the front page, but his friend Mary would see a mix of sports and business articles. He might vote on the articles he finds especially interesting or uninteresting, and these votes could be aggregated statistically with the votes of others in order to make

[39] New York Times, “Privacy Policy” (July 1, 2009), http://www.nytimes.com/ref/membercenter/help/privacy.html accessed 24 November 2009

recommendations about other articles he might like. In other words, in the same way that Netflix and Amazon recommend books or movies based on what users have read or seen, the *Times* might do likewise with articles. However, this might mean that it would have to bring in content from other sources. *Times* content alone might not be rich or diverse enough to create a truly individualized experience. It might be more interesting to analyze individuals' reading habits across several newspapers and blogs, than their reading on a single newspaper. Readers might identify collections of articles from multiple sources that they find to be related or interesting when read together. The resulting experience might be part blog, part Facebook conversational environment. Social bookmarking and filtering tools such as Delicious, Reddit and Digg make a pivot in this direction, although it is not clear if they are attempting to monetize individual behavioral activity yet.

Note that both of these projects, the vertical ad network and the integrated information experience, require publishers who might normally see themselves as competitors to partner with each other, and find some way to share ad revenues. Thus the logic of behavioral targeting pushes publishers towards forming their own aggregators. It does not represent a challenge to the argument that aggregators will become privileged relative to individual publishers; it just highlights how different kinds of aggregators can develop, offering different kinds of value propositions.

5.4 Ad networks and exchanges

Advertising networks and exchanges are currently the aggregators doing the most data collection. The exchanges might in fact change the ways in which networks differentiate themselves and the kinds of targeting they offer.

The exchanges force networks to compete more directly against each other, by making it easier for advertisers and publishers to switch between networks. Exchanges also force some level of standardization of the features networks provide. For example, if networks offer category-based behavioral targeting, it might be harder for each network to offer its own distinct set of categories in an exchange. Unstructured behavioral targeting, on the other hand, would encounter no such problem, because it would operate internally to the network. Thus the exchange model might push towards a one size-fits-all kind of categorical targeting, or else just unstructured targeting.

Exchanges might also change the data that is available to networks. As noted earlier, the exchange knows about the business relationships between publishers and networks: it knows, for example, that publisher P sells inventory through networks A, D and G but not networks B and C. Thus when a user visits P's website, the exchange learns about the user's visit and informs networks A, D and G, who in turn bid for the right to show an ad to the user. The exchange is a kind of "broadcaster" of the ad impression: A, D and G now all know about the user's visit to the site. Even if A and G lose the auction, the exchange still allows them to observe the user's behavior. Therefore the exchanges may make more information available to the networks than they would have had before. The only piece of information that is available to just one of the networks is the information about whether or not the user clicked on the ad. If D places the ad, only D knows whether the user clicked on the ad.

Thus the exchanges raise again the question of the relative value of behavioral data and ad inventory. If behavioral data is valuable to networks A and G, they may not care so much about losing the auction, and on balance the exchange helps them. But if they only care about ad placement, then the exchange hurts them because it forces them to compete more directly against other networks for the right to place ads. In short, exchanges may increase the number of parties that have access to behavioral data. This would in turn dilute the value of the data, if everybody's analysis of the data was the same. However, it could also force more innovation and competition with regards to the data analysis.

Another option for networks is to partner more closely with individual publishers, in order to make use of the more detailed profiles that publishers have available, integrating that data into the networks' behavioral targeting. They might try to make exclusivity agreements such that a publisher could only work with the one network and not share data with others, if the publisher data is valuable. The proliferation of vertical ad networks suggests that there are some advantages from specializing in particular topical areas; in other words, that a more focused domain-specific analysis of user behavior yields more useful insights than a non-vertical network could produce.

Finally, there is the question of whether the exchange itself is a threat to the networks. Since the exchange is connecting a great number of publishers with a great number of networks and advertisers, it also can collect an extensive amount of data about how the different networks target users, how users respond to the different targeting strategies of the networks, and what advertisers, agencies, and networks are willing to pay for inventory. It is conceivable that the company owning the exchange could use this data to build or aid its own network. The only thing that might prevent this is the exchange's desire to allow an ecosystem to develop. The exchange might want to encourage a large number of networks or partners to participate in order to make the overall market bigger. As long as no individual network gains enough market power to threaten the exchange, the exchange is happy to be a mediator between a large collection of small networks.

5.5 ISPs

ISPs are looking for a way to capture more value from their content. They have a challenge of become more than just "bit pipes", providing some extra value, whether to users by offering different tiers of content, or to content providers by providing prioritized service, or to advertisers by providing additional data. Content access is a lucrative business; the 2008 cable subscription revenues for Time Warner, just one of several large cable operators in the US, were $16 billion (Time Warner Inc., 2009), a number comparable in magnitude to the total amount spent on online advertising ($23 billion). However, the cable access market is nearing saturation, so it is conceivable that ISPs would be interested in tapping other revenue streams, such as the advertising market.

One advantage of ISPs compared to all of the existing advertising networks, publishers and ad agencies is that they have a complete view of their users' traffic (with the exception of encrypted communications, for example when users are managing financial accounts or making credit card payments online). All of the existing networks have "holes" in their view of Internet activity –

particular websites (or possibly whole categories of websites) that they do not monitor.

At a general level, network providers are in a good position to provide data on general traffic patterns: how many people from particular geographical areas are accessing a site at a given time. They could potentially be a credible independent producer of general audience measurements, such as the number of unique users visiting websites. The company Hitwise claims to produce such statistics by analyzing traffic data from ISPs. What is more difficult for ISPs to do is sell data about individual users. As noted earlier, several American ISPs have tried to do behavioral advertising, but encountered legal pressure, were forced to testify in Congress, and later retreated on these plans. The legal scholar Paul Ohm (2009) argues that current US law, in particular the Electronic Communications Privacy Act of 1986, could potentially forbid ISPs from monitoring the contents of their traffic, except at the most basic level or in situations where the ISP's property is under threat. However, the law itself is not clear, and this area of the law has not been heavily litigated, so there is some uncertainty about how the courts might actually rule.

The core question is how deeply ISPs are allowed to look into the contents of the packets they are carrying – and this is where the behavioral advertising issue connects with other regulatory and policy issues, specifically network neutrality and the use of deep packet inspection to detect illegal distribution of copyrighted content. Both of these debates are also about the extent to which ISPs can examine the traffic moving through their networks. At one extreme, some argue that the ISPs should only be allowed to examine the bare minimum of information necessary to route packets: the source and destination IP addresses. At the other extreme, some ISPs would like the ability to offer enhanced services to content providers for particular kinds of content (for example, more reliable streaming video). For example, a video hosting site like YouTube might pay Comcast (an ISP) an extra premium to deliver YouTube traffic in a faster or more reliable manner than other traffic. The FCC is currently investigating this issue with the goal of producing a set of guidelines about what kind of ISP traffic monitoring and treatment will be acceptable and not acceptable. These guidelines may also influence the extent to which ISPs can collect and sell data about users for advertising purposes. There is no legal prohibition against ISPs (or any other kind of player in this ecosystem) selling user data, only against monitoring.

If a strict kind of network neutrality were imposed, prohibiting ISPs from monitoring the contents of packets (in other words, prohibiting "deep-packet inspection" as discussed in Chapter 4), ISPs would still have be able to examine the source and destination IP addresses, which would provide some information. For all but the smallest websites, the IP address would identify which website the household is connecting to, but not the specific webpage. Thus the ISP could observe that a particular household connects to YouTube, Hulu, Amazon or eBay, but not which videos were watched or which products were viewed. (This is analogous to a phone company recording the phone numbers that a household has called, but not the contents of the call.) The fact that a particular household often connects to YouTube may not be particularly interesting for marketers, because so many households do so, but if the household connects to a less well-known site (such as a specialty site selling very large men's shoes) then the information may become more valuable precisely because it puts the household in a niche category. Because the ISP knows the household's address, it could compile and sell mailing lists of households with particular interests. The mailing address would be more valuable than the household's IP address because a household's IP address can change, rendering it unreliable for targeting purposes.

On the other hand, if a looser form of network neutrality were imposed, allowing the ISPs to look deeper into the content of their users' traffic, they might be allowed to collect information about which pages users were accessing. Thus network neutrality rules could, in the end, have the side effect of legitimizing certain kinds of ISP behavioral profiling.

Another option for ISPs is to get users' consent to behavioral tracking, for example by offering users a cheaper service if users allow the ISP to track their browsing activity. There are actually two models the ISP could use here. One is for the ISP simply to sell data about individuals' browsing patterns to other ad networks, ad agencies or anyone who is interested. Another is for the ISP to operate its own ad network, using the data it collects. The latter would be larger undertaking, requiring the ISPs to make deals with ad agencies, advertisers and publishers, but is still imaginable. (There have already been cases of ISPs replacing the "real" ads in a website with ads of the ISP's choosing, but it seems like such activity is not sustainable in the long term without buy-in from publishers.)

Yet another option is for ISPs to lobby directly to change the law to allow behavioral tracking, but it seems that the existing ad networks (including Google and Yahoo) would oppose them, potentially mobilizing public outrage against the spectre of ISPs trying to invade users' privacy. However, if the ISPs could offer a compelling case about the value of their behavioral advertising solution to enough advertisers and/or large content providers, they might be able to push through such a change.

It is also interesting to note that ISPs are pursuing a kind of targeted advertising in another context already: cable TV. As mentioned earlier, Project Canoe is attempting to develop a unified way for advertisers to target ads through cable networks to specific geographical areas or demographic segments. It also aims to enable interactive advertising, where for example the viewer could click a button to request more information about a product. This will make the TV experience more like an online experience. It also represents a way to capture more information about viewers. The average person in the US still spends much more time on the TV than online (Nielsen, 2009a), and the TV experience is seen as more of a passive experience where the user may be more receptive to advertising. This might explain why there are still more ad dollars in network and cable TV than the Internet, suggesting why it makes sense for the ISPs to focus on TV rather than online advertising.

In short, the question of ISPs' involvement in behavioral advertising has no simple answer, as there are several issues up in the air.

5.6 Users

There are three dimensions to this issue: users' norms, understanding and awareness of privacy as an issue; technological options (e.g. privacy-enabling technologies) that allow people to conceal their data; and finally, scenarios where users can exert more control over the aggregation of data.

On the one hand, online platforms encourage sharing of data with ever larger circles of friends. On Facebook, a note posted on one's page is instantly viewable by all of one's online friends, a much larger circle than would normally be possible to reach. Many of these "friends" may not be close friends in the normal sense of the word. So online activities may encourage a new level of sociality and a willingness to share information with others – the question is, will this willingness to share be extended to companies and marketers? In one sense, the Internet blurs the line between marketing and conversation. Users provide non-monetary value to marketers by discussing and recommending products to friends (Krauskopf, 2009). As people come to realize the value they provide to brands by discussing their products, they might become more resistant to traditional advertising, or come to ask for compensation for their services.

A recent telephone survey found that a majority of people were opposed to marketers tailoring advertisements to their interests, especially if it is based on their behavior on other websites or offline, and even if it takes places anonymously (Turow, King, Hoofnagle, Bleakley, & Hennessy, 2009). A majority also supported laws that would give people the right to see information that websites have collected about them and to request that such information be deleted. It is difficult for individuals to learn about how data about themselves is being used and sold. From direct mail one has some sense of who has bought a name from a mailing list, but not who sold the name. Likewise with the online environment, it is not clear who is observing behavior when visiting a page. A company could be tracking activity without even displaying ads, just by using a web bug.

As discussed in the previous chapter, there are tools available for users which block tracking cookies and ads from most well-known ad networks. (That said, many tools that claim to be "anti-spyware" are themselves spyware, which makes it more difficult for users to trust websites' claims about privacy (Zittrain, 2006).) One might wonder what would happen if a large share of Internet users were to become aware of these tools. This would have dramatic consequences for the online ecosystem, as online advertising spending would decrease and force content providers to find other ways to make money. Or, publishers might prohibit access to the content unless the user consented to tracking.

Thus while users may appear to have some control over whether or not they are tracked, publishers could, if necessary, design their sites so that tracking is required in order to access content. However, if users were sufficiently unhappy with this arrangement, they might find ways to share the content with each other, like people already do with copyrighted music and movies. Alternatively, if users do consent to tracking, they must to some extent trust the service provider's handling of your personal information. Technology can not guarantee this trust; it is, by necessity, social trust. Ideally, one would hope that privacy policies would become clearer about how data might be used or sold. There are cases where users have protested when online services did something that they felt violated the users' privacy – for example, when Facebook started sharing information about users' activities on other websites.

5.6.1 User choice or involvement in online tracking: some alternative scenarios

There are ways that users could be given more power, choice or at least knowledge about how their behavior is tracked. Several scenarios are imaginable.

One can imagine publishers allowing users to choose what kind of advertising they would like to receive, by selecting from a list of categories. However, it seems unlikely that this would satisfy the desire of technologists to make use of data as much as possible to do advanced prediction. The categories offered would have to be relatively broad, general categories; it is doubtful that users would be interested in choosing from a long list of very specific categories. On the other hand, just indicating an interest in "health" or "sports" is unlikely to be very useful for advertisers. Even if users were willing to choose from a greater number of categories, they might be reluctant to repeat this complicated selection process for many different publishers. There might be a desire to create a kind of shared profile that could be used by many different publishers. This would also make it easier for advertisers, because rather than having to deal with a variety of different taxonomies from different publishers, they would only have to deal with a small number.

Such a system might disrupt the process by which publishers sell large amounts of inventory in bulk directly to advertisers. A publisher could no longer guarantee showing one ad to all of its users, because they would each have different preferences, unless there was somehow a clear indication to the user that their preferences affect some ads but not others.

As a step in this direction, there are behavioral targeting systems that allow users to see the categories of content in which the user is believed to be interested. For example, Google's Ads Preferences Manager[40] displays the set of categories of sites on which Google has observed the user. While this can be lauded as a step forward in terms of giving users more awareness of what is happening, it is limited. It does not depict the full depth and detail of the data that Google has collected and finds of interest. It may indicate simply that a user is interested in "Travel" but in reality, the Google system may know many more specifics about what kinds of flights she has examined. Google may know that she searched for flights to Amsterdam last Thursday. If the full detail of data was revealed, people might be more concerned.

An alternative way to give the user power is to allow him or her to choose how his or her behavior will be tracked, in other words, allowing the user to choose from a set of behavioral tracking service providers. This would put the service provider in a more direct relationship with the users, rather than having the publisher take the first responsibility for the tracking relationship. Potentially the user could pay more or less depending on the amount of tracking that was done. Users could then opt for tracking only on certain categories of sites.

Another option is to use software running on users' computers to decide which ads to show based on the user browsing history stored on the computer. Ads would still be targeted, but there would be no need for a centralized database storing the browsing histories of all users. Toubiana et al. (Toubiana, Narayanan, Boneh, Nissenbaum, & Barocas) propose one way this might be done with a browser plug-in that downloads a set of candidate advertisements and then chooses from that list based on the user's browsing history.

Finally, there could be a way for users to "vote" on particular advertisements, without having to click on them and be interrupted from their current activity. This could be done with a set of checkboxes next to each ad that allow users to indicate which advertisements do and do not

[40] http://www.google.com/ads/preferences/

interest them. This would be a public admission of what has until now been a not-so-obvious fact: that an online advertising campaign is also a real-time survey, a means of gathering data about people's interests and desires. It would provide another kind of data to be used for targeting, while at the same time giving the user a feeling that she can contribute to the targeting process, rather than being a passive subject of observation. Of course, this would only work if the voting data was used instead of (and not in addition to) behavioral browsing data for ad targeting. Publishers might need to provide some rewards to users for voting, in order to guarantee that a sufficient amount of data is collected.

To make this system work, there would have to be enough advertisers or enough distinct advertisements in the system that users could actually be presented with different offers depending on their choices. Otherwise, users would get frustrated that their choices are being ignored. The risk to publishers (and potentially networks), of course, is that the advertisers that pay the most will be voted down by users.

One might ask if the user would have any incentive to misreport or lie about his interests. If the ads dealt with sensitive personal materials (such as health or finances) a user might be reluctant to share an opinion. If the user really hates advertising, he might enjoy entering false or misleading information into the system – a kind of "click fraud". Related to this, a company could hire people to vote down the offers from its competitors, potentially making the competitors' ads less likely to appear. A voting system could be manipulated in different ways than current advertising systems, because current systems only allow two kinds of responses to an ad (clicking and not clicking), while a voting system has four (positive vote, negative vote, clicking, and no action).

All of the options described in this section would require cooperation from publishers and potentially advertising networks. Users could not implement these systems on their own. The question is whether they could still provide enough revenue from the kind of targeting they provide. It would also be desirable, with any of these systems, to allow third-parties to audit the collection of data to the extent possible (for example, by still using cookies and JavaScript so that activity could be observed). In addition, many of these systems might be more effective with a strong form of user identity that persists across sites and does not easily disappear, like cookies currently do. Alternatively, users could create accounts on the third-party tracking system and then provide that username to publishers.

In short, there are a number of conceivable ways to achieve some kind of middle ground between the current form of behavioral targeting (which is surreptitious and not obvious to users) and a blanket ban on aggregation of personal data. However, the value propositions and business models for these targeting systems still need to be elaborated.

5.7 Conclusion

The analysis here points to several ways in which the drive for behavioral data is helping aggregators to develop, be they publishers, agencies, networks, or (though it seems less likely) ISPs. If they are aggressive enough, the agenices might gain a stronger position by working with publishers more directly and building their own networks. But if agencies do not move in this manner, it appears that networks and exchanges will be in the strongest position. For publishers to gain ground, they would have to make a dramatic shift, such that they would no longer be "publishers" in the true sense of the word, but rather "information experiences". It is conceivable that users could gain more control over behavioral tracking, but this would require more willingness by networks and advertisers to engage with users' privacy concerns.

This thesis will close with several additional general thoughts about the evolution of online advertising.

Advertising agencies are still adapting to the possibilities and challenges of the Internet. They are caught in varying degrees between the past and the future. In a sense, they represent a kind of "inertia" from the pre-Internet era: the old model of buying mass media (TV, cable, print and newspaper) at large scale remains. It will take some more time yet before new models stabilize.

Behavioral targeting challenges big brands' traditional strategy of broadcasting one message to millions of people, and may be more suitable for small-scale marketers. More generally, the Internet may force changes in the entire concept of "branding". As discussed earlier, behavioral targeting is presently not a tool for brands interested in reaching a mass audience. The total costs of behavioral targeting campaigns can be less than mass media buys (even though CPMs might be higher), making it easier for small marketers to use behavioral targeting. Big brands may be hurt because they are forced to bid their CPMs up in competition with the niche advertisers who are willing to pay higher CPMs. Behavioral targeting might then have a different base of political support to lobby against privacy regulations.

Before the Internet, "branding" largely meant 30-second TV spots, print ads, and product packaging. Advertisers could control the brand because there were limited ways for users (consumers) to communicate with each other and tell their own stories about products. With the Internet, however, users have more control over products' social meanings.

Behavioral data collection will merge with other kinds of market research, and behavioral targeting may merge with social media marketing. The technologies are still evolving and the payoffs from various technologies are unclear. It is tempting to look at the future through a couple of different scenarios: one where behavioral targeting is wildly successful (for marketers) and another where it fails. Rather the future is likely to be more complex, where behavioral targeting merges with social media marketing and other kinds of market research[41]. These are all a variety of projects attempting to classify and categorize people based on their online activity. "Behavioral targeting" is currently focused on just the sequence of web pages that a user visits, but in the future, it might be combined with peoples' contributions to social media sites, their social graph and offline activity, such as TV watching and in-store shopping. Agencies might be in the best position to do this kind of integrated research. It is still not clear that behavioral targeting by itself will yield huge improvements in marketing ROI, but behavioral data might have value for market research.

It does not seem likely that behavioral targeting itself would lead to a significant overall

[41] For one agency's take on this, see Razorfish (2009).

increase in spending on online advertising. There are other factors restraining the growth of online advertising. As discussed earlier, unless a large portion of publisher inventory can be targeted at high rates, the overall revenue received by publishers will not increase significantly. Thus behavioral targeting is not itself a solution to publishers' woes. Furthermore, it makes publishers' task more complex as their revenue becomes determined by user data collected by third parties completely outside the control of publishers. If ad targeting is based primarily on user profiles, prestigious publishers' content has less weight in the marketplace. Why would advertisers pay extra to reach a user on the *New York Times* if they can reach him on other sites for much less? (Of course, this assumes a buyers' market, where there is an abundance of inventory and places to reach users, and a relative scarcity of advertisers willing to pay premium prices.) It is hard enough for publishers to determine the demographics of their audiences online. It is harder still for them to know about their users' interests, unless they become much more interactive and increase their "depth of identity".

If there is to be significant growth in online advertising spending, it will be the result of a number of changes, including advertisers and agencies gaining more experience with online advertising. A new concept of "branding" may be part of this, as will the development of standard metrics for audience measurement (Nielsen, 2009b). The web offers many different kinds of measurements, making it difficult for the ecosystem to settle on one particular model. It is not that the current online metrics are necessarily worse than their counterparts in other media; Nielsen TV surveys had problems of their own, but people accepted them because they were "good enough" and because there were no other options. On the Internet, however, any number of metrics have been proposed, and no clear winner has emerged. Potentially, after enough experimentation, some combination of actors with sufficient weight will settle on a set of standardized (though imperfect) metrics. Ad exchanges might be a driving force.

People have different privacy expectations in different contexts; an explicit acknowledgment of this fact might help assuage some of the concerns. The privacy expectations for web mail might be higher than for a public discussion board. The privacy expectations for Facebook are currently being negotiated. Users will become more aware about privacy issues and in turn come to demand more transparency from service providers about the level of privacy they provide. Potentially some of the privacy concerns could be addressed by an explicit recognition that there are different kinds of online "privacy environments" where different kinds of privacy norms apply[42]. A "privacy environment" could be defined as a collection of affiliated sites that share data with each other about user behavior and track users as they move between the member sites. Data does not move from one privacy environment to another, and users would have different IDs in different environments so their activities could not be cross-referenced. Each privacy environment would have a logo that is displayed in an obvious place on all of the member sites (perhaps near the ads); clicking on the logo would present a page with some explanation about how data is shared between the sites. A further elaboration of this idea would be for the privacy environment to give the user some further choice or control over the kind of tracking, as discussed in the previous section.

Branded vertical networks, i.e. ad networks operated by a brand-name publisher like Fox or MTV but also serving ads on a hand-picked collection of lesser-known sites, could be one kind

[42] Thanks to Dan Pereira at the MIT Convergence Culture Consortium for suggesting this idea.

of privacy environment, assuming that they do not share data with external parties. On the other hand, ad exchanges might facilitate the exchange of user data and cross-referencing of identities between different actors and environments, and thus threaten the sharp delineation of such environments.

Related to the previous point, different modes of data collection are more or less obvious to users; if data collection is to be accepted, perhaps it is better to make it more obvious, and/or give users some choice or control over the process. "Recommendation engines" are good models of how the user benefits of behavioral targeting can be achieved in ways that are also more transparent to users. Like a recommendation engine on a site such as Amazon.com, a behavioral targeting system suggests other products a user might be interested in based upon his past choices and potentially what he looked at in the past but did not buy. Unlike Amazon's recommendation engine (it is believed), the placement of offers in front of a person is also partly determined by how much advertisers are willing to pay. A recommendation engine can also be based on the recommendations or actions of friends, for example, social network connections, or user votes. But in all cases, it should be clear to users that the recommendations are based on past activity or behavior, which is not the case with current behavioral targeting systems.

If Facebook ads become increasingly based on the contents of individuals' social expression, users may become more aware of how their data can be used for marketing purposes, and the privacy tradeoffs associated with different online services. Behavioral targeting can also be obvious, for example in the form of re-targeting, which reminds a user about a product they were recently viewing but did not purchase. But it can also be more subtle, and therefore harder for the user to notice. One might argue that the more obvious the profiling, the better, because it makes clear to users the bargain they are making with online services. Subtle and complex profiling is less likely to be noticed by consumers and therefore more insidious. Of course, marketers may be reluctant to make their profiling too obvious; this is the "cat and mouse" game that advertisers play.

As discussed in the previous section, users could be offered more choice or control of targeting, which again would require marketers to cede some control over how advertising is distributed – in short, making advertising more like other forms of content.

A distributed open-source project could gather useful data for researchers and policymakers about the extent of aggregators' observation of users, as well as about online advertising in general. This might build from the work of Krishnamurthy and Wills (2009) discussed earlier, but access a greater variety of pages on different sites, and from a variety of locations and potentially with a variety of cookies so that it could measure the extent to which different ads are shown to different users. The greater the number of people involved in the project, the more it could observe subtle forms of targeting. If only a few people were involved in such a project, it would be statistically difficult to know the extent to which the ads seen by one person were personalized. Such a project would be analogous to a number of other distributed data-collection projects, including Herdict[43], which attempts to collect data about Internet censorship in a distributed manner.

[43] http://www.herdict.org/web/

This leverages the open and generative nature of the web, taking advantage of the web’s ability to “observe itself”. The current mode of behavioral tracking, using browser interactions, Javascript and cookies, makes it in some sense “public”, at least to technically savvy users. This differs from the buying and selling of mailing lists, which is harder to “reverse engineer”. Of course, the proposed project could not observe how data from publishers and networks is shared behind the scenes, i.e. directly between servers, rather than through browser-based mechanisms. In addition, such a project would have privacy concerns of its own, as it would involve the aggregation of browsing data from a number of computers.

Bibliography

Abraham, M., Meierhoefer, C., & Lipsman, A. (2007). *The impact of cookie deletion on the accuracy of site-server and ad-server metrics: An empirical ComScore study.* Retrieved from http://www.comscore.com/Press_Events/Presentations_Whitepapers/2007/Cookie_Deletion_Whitepaper

Advertising Age. (2009). *Ad Network+Exchanges guide.* http://brandedcontent.adage.com/adnetworkguide09/

Baker, S. (2008). *The numerati*. Boston: Houghton Mifflin Co.

Bermejo, F. (2007). *The internet audience: Constitution & measurement*. New York: Peter Lang.

Blumenthal, M. S., & Clark, D. D. (2001). Rethinking the design of the internet: The end-to-end arguments vs. the brave new world. *ACM Transactions on Internet Technology, 1*(1), 70-109.

Broder, A., & Josifovski, V. (2009). *Introduction to computational advertising.* Retrieved 9 December, 2009, from http://www.stanford.edu/class/msande239/

Cappo, J. (2003). *The future of advertising : New media, new clients, new consumers in the post-television age*. Chicago: McGraw-Hill.

comScore. (2009). *The comScore 2008 digital year in review.* Retrieved from http://www.comscore.com/Press_Events/Presentations_Whitepapers/2009/2008_Digital_Year_in_Review

DeSilva + Phillips. (2008). *Online ad networks: Monetizing the long tail.* Retrieved from http://www.iab.net/media/file/AdNetworksWhitePaper.pdf

Direct Marketing Association. (2009). *Statistical fact book 2009.* New York: Direct Marketing Association.

Downie Jr., L., & Schudson, M. (2009). *The reconstruction of american journalism.* Retrieved from http://www.journalism.columbia.edu/cs/ContentServer?pagename=JRN/Render/DocURL&binaryid=1212611716626

Edelman, B. (2009). *Towards a bill of rights for online advertisers.* http://www.benedelman.org/advertisersrights/

Hallerman, D. (2008). *Behavioral targeting: Marketing trends.* eMarketer. Retrieved from http://totalaccess.emarketer.com/GetFile.aspx?type=re&code=emarketer_2000487

Hu, J., Zeng, H., Li, H., Niu, C., & Chen, Z. (2007). Demographic prediction based on user's browsing behavior. Paper presented at the *WWW '07: 16th International Conference on World Wide Web,* Banff, Alberta, Canada. 151-160. Retrieved from http://doi.acm.org/10.1145/1242572.1242594

Interactive Advertising Bureau, & PricewaterhouseCoopers. (2009). *IAB internet advertising revenue report.* Retrieved from http://www.iab.net/media/file/IAB_PwC_2008_full_year.pdf

Jenkins, H. (2006). *Convergence culture: Where old and new media collide.* New York and London: NYU Press.

Karpinski, R. (2009a, April 20). 'Demand-side' networks give agencies a boost in display. *Advertising Age*

Karpinski, R. (2009b, April 20). What to expect from the next gen of digital display. *Advertising Age*

Krauskopf, A. D. (2009). *Tacky and proud: Exploring tecnobrega's value network.* MIT Convergence Culture Consortium.

Krishnamurthy, B., & Wills, C. E. (2009). Privacy diffusion on the web: A longitudinal perspective. Paper presented at the *World Wide Web Conference,* Madrid, Spain. Retrieved from http://www.research.att.com/~bala/papers/www09.pdf

McClellan, S. (2009, May 30). IPG unveils cadreon digital ad network. *AdWeek*

Nielsen. (2009a). *A2/M2 three screen report, 3rd quarter 2009.* Retrieved from http://en-us.nielsen.com/main/insights/reports

Nielsen. (2009b). *Building great brands in the digital age: Guidelines for developing winning strategies.* Retrieved from http://en-us.nielsen.com/forms/report_forms/building_great_brands

Ohm, P. (2009). The rise and fall of invasive ISP surveillance. *University of Illinois Law Review, 2009*(5), 1417.

O'Leary, N. (2009, May 25). Searching for life on hulu. *AdWeek*

Razorfish. (2009). *Razorfish digital outlook report 09.* Retrieved from http://www.razorfish.com/#/ideas/reports-and-papers/special-reports/

Soltani, A., Canty, S., Mayo, Q., Thomas, L., & Hoofnagle, C. J. (2009). *Flash cookies and privacy.* Retrieved from http://ssrn.com/paper=1446862

Spangler, T. (2009, June 18). EXCLUSIVE: Canoe scraps initial zone-ad plans. *Multichannel News,* Retrieved from http://www.multichannel.com/article/295253-EXCLUSIVE_Canoe_Scraps_Initial_Zone_Ad_Plans.php

Spurgeon, C. (2008). *Advertising and new media*. London and New York: Routledge.

Time Warner Inc. (2009). *Annual report 2008.*

Toubiana, V., Narayanan, A., Boneh, D., Nissenbaum, H. & Barocas, S. *Adnostic: Privacy preserving targeted advertising.* http://crypto.stanford.edu/adnostic/

Turow, J. (2006). *Niche envy : Marketing discrimination in the digital age*. Cambridge, Mass.: MIT Press.

Turow, J., King, J., Hoofnagle, C. J., Bleakley, A., & Hennessy, M. (2009). *Americans reject tailored advertising and three activities that enable it.* Retrieved from http://ssrn.com/paper=1478214

US Federal Trade Commission. (2007). *Statement of federal trade commission concerning Google/DoubleClick, FTC file no. 071-0170.* Retrieved from http://www.ftc.gov/os/caselist/0710170/071220statement.pdf

Wetpaint, & Altimeter Group. (2009). *The world's most valuable brands. who's most engaged?* Retrieved from http://www.engagementdb.com/downloads/ENGAGEMENTdb_Report_2009.pdf

Winterberry Group. (2009). *The data-driven web: Targeting, optimization and the evolution of online display advertising.* Retrieved from http://www.winterberrygroup.com/ourinsights/wp

WPP. (2009a). *Google - friend or froe? - WPP annual report & accounts 2008.* http://www.wpp.com/annualreports/2008/what_we_think/insight/google.html

WPP. (2009b). *Google and WPP marketing research awards program bestows 11 grants.* Retrieved August 5, 2009, from http://www.wpp.com/wpp/press/press/default.htm?guid={E0AF399A-8450-408C-8BA8-C35D31DAE88C}

Zittrain, J. L. (2006). The generative internet. *Harvard Law Review, 119,* 1974.