Challenges in Measuring Online Advertising Systems

Saikat Guha
Microsoft Research India
Bangalore, India
saikat@microsoft.com

Bin Cheng, Paul Francis
Max Planck Institute for Software Systems
Kaiserslautern-Saarbruecken, Germany
{bcheng,francis}@mpi-sws.org

ABSTRACT

Online advertising supports many Internet services, such as search, email, and social networks. At the same time, there are widespread concerns about the privacy loss associated with user targeting. Yet, very little is publicly known about how ad networks operate, especially with regard to how they use user information to target users. This paper takes a first principled look at measurement methodologies for ad networks. It proposes new metrics that are robust to the high levels of noise inherent in ad distribution, identifies measurement pitfalls and artifacts, and provides mitigation strategies. It also presents an analysis of how three different classes of advertising — search, contextual, and social networks, use user profile information today.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement techniques; K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*

General Terms

Experimentation, Measurement

Keywords

Advertising, Privacy, Behavioral Targeting, Contextual, Churn, Similarity, Google, Facebook

1. INTRODUCTION

Online advertising is a key economic driver in the Internet economy, funding a wide variety of websites and services. At the same time, ad networks gather a great deal of user information, for instance users' search histories, web browsing behaviors, online social networking profiles, and mobile locations [6–8]. As a result, there are widespread concerns about loss of user privacy. In spite of all this, very little is publicly known about how ad networks use user information to target ads to users. For instance, Google recently started allowing advertisers to target ads based not just on keywords and demographics, but on user interests as well [9]. Knowing how well Google and others are able to determine user characteristics is an important consideration in the ongoing public debate about user privacy.

If an ad network is able to accurately target users, we can deduce that the ad network is able to determine user characteristics (though the inverse does not follow). Given then the goal of determining how well ad networks can target users, the high-level methodology is straightforward. Create two clients that emulate different values of a given user characteristic (i.e. location or gender), and then measure whether the two clients receive different sets of ads as a result. If the ads are identical, we can trivially conclude the ad network doesn't use that user characteristic for targeting (although it might still be storing the data). But the outcome is unclear if the two sets are different: the difference may genuinely be due to the difference in characteristic, or it may be due to noise.

As it turns out, the level of noise in measuring ads is extremely high. Even queries launched simultaneously from two identically configured clients on the same subnet can produce wildly different ads over multiple timescales. As we show later, some of this noise is systemic (e.g. DNS load-balancing), and can therefore be eliminated through proper experiment design. Other noise has a temporal component, which likely reflects ad churn, the constant process of old ads being deactivated and new ads being activated. We design a metric that mitigates the noise from churn.

Overall this paper makes two contributions. First, we present the detailed design of a measurement methodology for measuring online advertising that is robust to the high levels of noise inherent in today's systems. We present a set of guidelines for researchers that wish to study advertising systems. Second, we present an analysis of the key factors that determine ad targeting on Google and on Facebook.

2. MEASUREMENT METHODOLOGY

In this section we present the detailed design of our measurement methodology. We face four key challenges: 1) comparing individual ads, 2) collecting a representative snapshot of ads, 3) quantifying differences between snapshots while being robust to noise, and 4) avoiding measurement artifacts arising from the experiment design.We present our design decisions and justify each using measurement data. The methodology we design applies to text-ads in any context including ads in search results, contextual ads on webpages

IMC'10, November 1–3, 2010, Melbourne, Australia.

Instance A: Red Prom Dresses Win a Free Dress for Prom 2010. Find New Trends; Great Prices! `DavidsProm.com` MD5(RedirURL): 8ebc. . . 45dc Dest: . . . detail.jsp?i=2462	**Instance B:** Red Prom Dresses Beautiful Designer Prom Dresses to Fit Every Figure; Price Range. `DavidsProm.com` MD5(RedirURL): 3646. . . 85d3 Dest: . . . detail.jsp?i=2462
Instance C: Baby Doll Prom Dresses Win a Free Dress for Prom 2010. Find New Trends; Great Prices! `DavidsProm.com` MD5(RedirURL): 99d0. . . f0bf Dest: . . . detail.jsp?i=2203	**Instance D:** Red Prom Dresses Shop JCPenney For Colorful Prom Gowns; Dresses From Top Designers. `JCPenney.com/dresses` MD5(RedirURL): c12d. . . ce2c Dest: . . . X6.aspx?ItemId=17bd2fb

Table 1: Examples that complicate uniquely identifying ads.

and webmail systems, and ads on online social networking pages; additionally, many qualitative aspects of our design may also apply to banner ads.

2.1 Comparing Individual Ads

Before we can compare sets of ads, we first need the ability to identify the same ad in different scenarios. The problem is hard because ad networks typically do not reveal the unique ID for the ad (except for Facebook), and some (like Bing) even go so far as to obfuscate or encrypt parameters in the click URL denying scraping based measurement approaches any visibility into internal parameters. The only data available consistently across ad networks is the content of the ad, and even there the extensive abilities to customize it (e.g. for Google), makes it hard to identify different instances of the same ad. Since slight variations of the same ad defeat simple equality tests, heuristics must be used and their false positive and false negative behavior must be understood.

Consider, for example, the four ads illustrated in Table 1. Instances A and B are semantically equivalent, mutually exclusive (i.e. never both served for the same request), and lead to the same page on the advertiser's website (despite having different redirect URLs[1]). Instances A and C appear to be from the same template, but are semantically different and lead to different pages. Instances A and D are from different advertisers altogether.

Ideally, instances A and B would be considered equivalent, and different from both C and D. This cannot be achieved with simple equality tests on any single ad attribute (title, summary, display URL[1], redirect URL); other examples not presented here demonstrate the presence of false positives or false negatives for equality tests on combinations of attributes.

Experiment 1: Since comparison errors are unavoidable, we analyze false positives and false negatives of different approaches to comparing ads in order to pick the best approach as well as provide a bound on analysis errors. We consider 4 approaches: 1) equality of the redirect URL, 2) equality of the display URL, 3) equality of the ad title and display URL, and 4) equality of the ad title and summary text with all occurrences of the search keywords masked. Note: we cannot consider the destination URL for the ad since that is revealed only after the ad is clicked (and clicking on the ad in an automated matter constitutes fraud). We apply each approach to all pairs of ads in two datasets of ads scraped from Google search results. For estimating false positives, we manually check a sampling of pairs flagged as equal by the comparison approach. For false negatives, we manually check pairs flagged as different by the comparison approach; to focus manual analysis on pairs likely to be false negatives, we examine those flagged as equal by one of the other approaches, thus providing a lower bound. The first dataset contains ads for search queries related to fashion in general (clothing, shoes, accessories, etc.), while the second dataset restricts the queries to only dress-related.

	All Fashion		Dresses only	
Approach	% FP	% FN	% FP	% FN
RedirURL	0	38	1	52
DisplayURL	7	13	12	10
Title + DisplayURL	0	45	0	50
Title + Summary	0	68	0	69

Table 2: False positives (FP) and false negatives (FN) of various approaches for uniquely identifying ads.

Table 2 summarizes the false positives and false negatives for each comparison approach based on manual analysis of 100 ad-pairs flagged by each approach. Except when comparing display URLs, there are (almost) no false positives. The false positives for display URLs arise from some stores using the same display URL for a wide range of products (e.g. `target.com` instead of `target.com/mens`). The display URL does, however, have significantly lower false negatives in both datasets. This is because advertisers often have multiple variations of the same ad. This includes different title, summary, and a different redirect URL to measure how each variation performs; comparing ads based on these is therefore more error prone. The display URL, in contrast, is more stable across variations of the same ad.

False negatives have the effect of over-counting the number of unique ads and in the process, increasing noise. False positives, on the other hand do not increase noise, but under-count the number of unique ads. In this paper we perform all analysis relative to a control experiment (affected equally by false positives or negatives) to mitigate the effect of over-counting or under-counting. This leaves added noise (for false negatives) as the only differentiating factor. In this paper we therefore use the display URL, which has significantly lower false negatives and slightly higher false positives, to compute uniqueness of ads.

2.2 Taking a Snapshot

There typically are more ads than can be displayed on a single page (search result, or adbox in a webpage). A single query therefore reveals incomplete information. The actual subset returned depends on the ad network, but likely takes into account the ad value (based on auctions), and frequency capping (not showing the same ad to the same user too often). Reloading the page multiple times typically reveals more ads, but runs the risk of capturing inaccurate snapshots in the presence of ad churn.

Experiment 2: To determine how many times the page should be reloaded to capture accurate and complete snapshots we perform the following experiment. We reload the Google search results for a given query every 5 seconds for 5 minutes; the experiment is repeated for over 200 different

[1] Display URL (e.g. `DavidsProm.com`) is the URL displayed to the user. Redirect URL (RedirURL), is the URL the user is actually redirected to when he clicks the ad. The redirect URL is longer and less user-friendly, containing deep path information and URL parameters. The destination URL (Dest) where the user is eventually taken to may be different from the redirect URL when multiple redirects are involved.

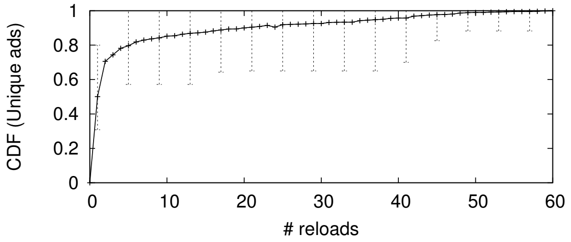


Figure 1: Reloading the page initially reveals more ads that didn't originally fit. Beyond a point, however, ad churn overtakes diminishing returns of reloading the page. 95-5 error bars.

search queries (chosen randomly from a set of 5000 keywords scraped from Google Product Search[2]).

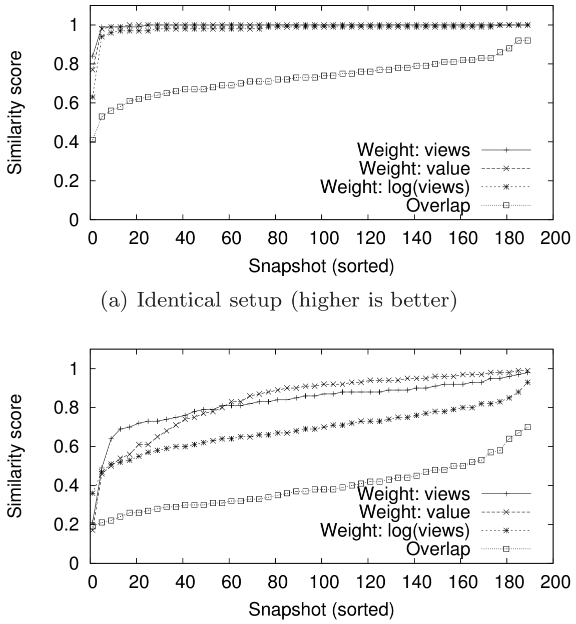
Figure 1 plots the CDF of unique ads for the median search query, with error bars marking off the 95^{th} and 5^{th} percentiles. The graph comprises a steep increase (5 reloads) where we discover most ads that didn't fit the first time, after which point diminishing return kicks in. Instead of flattening out, however, beyond 10 requests the graph becomes linear. This gradual linear increase, we believe, is evidence of a constant rate of ad churn where new ads are constantly activated and old ones deactivated. The high rate of churn (1–4% per minute depending on the search keyword) is consistent with [5] where we found that roughly only 60% of ads are stable (hour-to-hour, and day-to-day) while the rest change rapidly. The knee of the graph (around 10 reloads) represents the point at which the diminishing returns of reloading the page is overtaken by ad churn.

In our experiments we therefore balance completeness and churn by reloading the page 10 times when collecting snapshots. We also keep track of the number of times each ad was seen (if multiple reloads contain the ad) and the positions where the ad was seen.

2.3 Quantifying Change

The simplest approach to comparing two snapshots is to compute the set overlap. The Jaccard index quantifies this overlap as $\frac{|A \cap B|}{|A \cup B|}$ where A and B are the sets of unique ads in the two snapshots; 0 implies no overlap and 1 implies identical snapshots. While simple, the Jaccard index is highly susceptible to noise (fleeting ads); each unique ad is weighed equally whether it was seen only once or seen many times. The typical way of dealing with noise is to look at aggregate behavior. Taking the union of multiple snapshots, however, makes the situation worse by also aggregating the noise.

The extended Jaccard index (also called cosine similarity) addresses this limitation by interpreting the two sets as vectors in n-dimensional space (where each set element defines one dimension), and the coefficient of the vector in that dimension is some weight function (w) based on the element. The metric is defined as $\frac{\bar{A} \cdot \bar{B}}{\|\bar{A}\| \|\bar{B}\|}$ where $\bar{A} = [w_{A,e}]$; $w_{A,e}$ is some non-zero weight if ad e exists in A, or 0 if it doesn't. As before, the metric evaluates to 0 for dissimilar snapshots, and to 1 for identical snapshots. We explore three approaches to picking the non-zero weight: 1) the number of page reloads containing the ad, 2) the logarithm of the same, and 3) the number of page reloads scaled by the "value" of the ad; the value, defined in [4], is based on the ad's position with ads near the top of the page getting more weight than those near the bottom. By taking the number of times the ad was seen into account, each weight function also effectively attenuates noise when snapshots are aggregated together.



(a) Identical setup (higher is better)

(b) Different setup (lower is better)

Figure 2: CDF of similarity scores computed by the four metrics tracked over 8 days.

Experiment 3: To determine which approach performs the best we conducted the following experiment. We simultaneously collect snapshots from two browser instances configured identically and on the same machine (in New York); we expect the snapshots to be substantially the same. We also simultaneously collect snapshots from a machine set up at a remote location (in San Francisco) where we expect to see some differences in the set of ads. Snapshots (for 15 queries) are collected every 5 minutes for a period of 8 days. We then aggregate 1 hour's worth of data (12 snapshots) and compare the performance of all four metrics: the Jaccard index, and the extended Jaccard index with the three weight functions.

Figure 2(a) plots the CDF of the computed metric value for the case where we expect snapshots to be identical; the closer to $y = 1$ the better. As expected, the plain Jaccard index performs poorly. The other three perform quite well, with the logarithmic weight function trailing slightly due to the reduced influence of highly-stable ads. Figure 2(b) plots the CDF for the case where we expect snapshots to be different; the greater the difference between the corresponding lines between 2(a) and 2(b) the better. The logarithmic weight clearly outperforms the other two here. This is because for the other two weight functions, and especially so for weights based on the ad value, the long tail of ads is drowned out by a handful of highly-stable highly-ranked ads, which tend to be from large companies (e.g. eBay, Amazon) that target broadly across many demographics, interests, and locations.

In our analysis we use the extended Jaccard index with logarithmic weights to quantify the similarity or difference between snapshots. The metric in practice is both robust to

[2] `http://www.google.com/products`

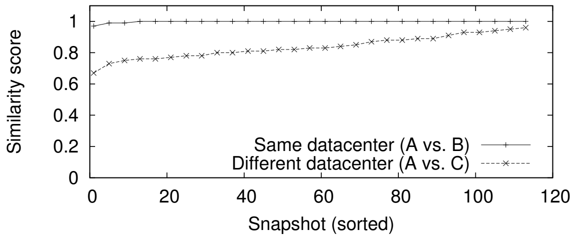


Figure 3: Artifact of DNS load-balancing by ad network.

noise, as well as sensitive to changes in the underlying set of ads.

2.4 Avoiding Artifacts

During the course of our experiments we identified several measurement artifacts that stemmed from poor interactions between lower level protocols and the operational architecture of the ad network. We discuss two artifacts and how they can be eliminated: the first pertains to DNS load-balancing, and the second to distributed data collection from multiple machines.

In one case we observed discrepancies between data collected by three identically configured browser instances all running on the same machine. Snapshots for the first two browser instances (A and B) are virtually identical, while that for the third instance (C) is very different (Figure 3). We discovered the ad network domain is DNS load-balanced and the browsers did not share a DNS cache (and therefore each queried DNS independently). Instance A and B were communicating with different IP addresses in the same /24 (same city), while C was communicating with an IP address in the same /16 but different /24 (different nearby city), which we take to mean two different datacenters. Thus even for identical requests from the same source, the choice of datacenter, we found, dramatically affects the ads served. This artifact is, of course, easily avoided by configuring a static entry in the hosts file so all instances reach the same datacenter.

Even with static DNS entries, we sometimes (but not always) observed discrepancies when running identical browser instances on different machines. The 5^{th} percentile similarity score dropped to 0.87 for different machines (compared to 0.99 in the same-machine case). We noticed the cookies assigned to the two instances were different; when we synchronized the cookie values, the noise disappeared. In another case, we measured similar levels of noise when we added a HTTP proxy in front of the two machines (which downgraded HTTP/1.1 to 1.0). We believe these artifacts are because of black-box frontend load-balancer behavior at the ad network that, based on at least the IP address, HTTP version, and cookies, we suspect, directs the requests to different backend servers, each of which has a slightly different cache of ads. The only way to mitigate this source of noise, we believe, is to ensure as many header fields are held constant as possible. Ideally, traces for an experiment are all collected from a single IP address, not behind a proxy, with cookies synchronized across browser instances.

Applying these techniques significantly reduces the base noise level, but does not eliminate it. We therefore measure the noise during our experiments to detect anomalies and to establish a level of confidence in our results.

3. ANALYSIS

In this section we use the above methodology to explore specific questions regarding how ads are targeted in three different contexts: search, websites, and online social networks. These questions include, among others, whether behavioral targeting affects search ads, whether past searches affect ads on websites, and what profile data affects social network ads. That said, since ad targeting is a black-box where we can reliably control only a small set of inputs, we are restricted in the questions we can answer. An example of a question we cannot answer is whether Google learns the user's gender by observing which search results the user clicks and then uses it to target ads; this is because we cannot reliably affect or verify the gender learned by Google's (black-box) algorithm if it indeed does so at all.

For questions where we can reliably affect the inputs to the ad selection algorithm, our experimental methodology is as follows. For each experiment we configure two (or more) measurement instances to differ by exactly one input parameter, and configure two measurement instances identically to serve as the noise-level control. If the similarity score between the control pair is high (i.e. low noise) but that between two differently configured instances is low, we conclude that the input parameter in which the two instances differ affects the choice of ads. For scalability and repeatability, all experiments are scripted using the Chickenfoot browser automation framework [2].

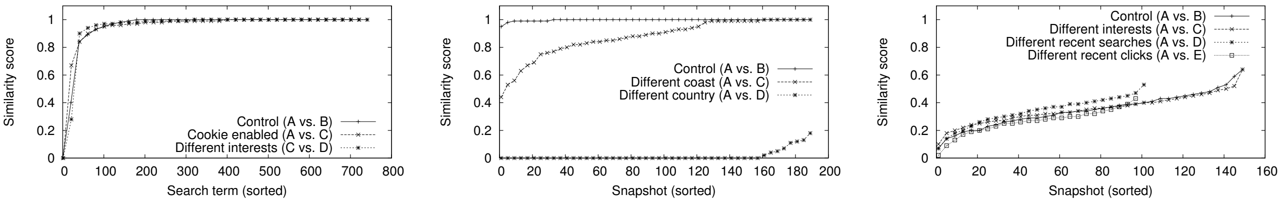
3.1 Search Ads

Search ads have typically been targeted based on keywords in the search query. It is therefore expected that keyword based targeting dominates search. The question we ask is: to what extent does behavioral targeting affect search ads? Behavioral targeting refers to using the user's browsing habits to influence ad selection.

Experiment 4: We set up four browser instances: the first two (A and B), which also serve as our control, disable DoubleClick's DART cookie [3] that Google uses for behavioral targeting. The third (C) and fourth (D) have cookies enabled, but are seeded with different user personae[3]. C was seeded with long-term interests in 'Autos & Vehicles', while D was seeded with interests in 'Shopping'. We then perform Google searches for 730 random product-related queries for a period of 5 days.

Figure 4(a) illustrates to what extent behavioral targeting affected search ads. As is evident from the figure, for keywords where the data is not too noisy (i.e. control score is high), there is no appreciable difference in the ads served whether the behavioral targeting cookie is disabled or enabled (A vs. C), or for two users with different interests (C vs. D). To understand why, we looked at the ads served. 73% of them contained the whole search query somewhere in the ad, and 97% of them contained at least one word from the search query. It is therefore clear that ads are selected primarily based on keywords. Furthermore the average number of unique ads for our search queries is 8. Since all the ads matching the keyword can be shown to the user in a

[3] Google normally learns short-term and long-term user interests completely automatically. It also allows users to view and modify the learned interests (`http://www.google.com/ads/preferences`), which we use to create (or verify Google learned) different personae.



(a) (b) (c)

Figure 4: (a) Behavioral targeting doesn't appear to affect search ads. (b) Location affects website ads. (c) The affect of user behavior (browsing, search, clicks) on website ads is indistinguishable from noise in the system.

short period, there may be little reason to pick and choose any further.

3.2 Website Ads

Website ads have typically been based on the context of the page and are hence also known as contextual advertising. We ask here to what extent the user's location affects the choice of ads, and to what extent the user's behavior (browsing behavior, recent searches, and recent clicks on products) affects website ads. We measure a set of 15 websites that show Google ads; the websites are picked randomly from the set of websites visited by CoDeeN [1] users. While we are able to answer the location question, we are able to present only weak evidence towards the lack of use of behavioral data due to noise.

Experiment 5: To understand the impact of user location, we set up four browsers: A and B, the control pair, in the same city in the US (New York), C in a different city on the other coast (San Francisco), and E in a different country (Germany).

Figure 4(b) plots the similarity scores between the different instances. As one might expect, location affects the set of ads, but interestingly, there is (relatively) little difference between cities on opposite coasts (median similarity of 0.9). This is higher than we expected given the long-tail of ads appears to contain local mom-and-pop retailers, although in retrospect, these retailers may nevertheless conduct business nation wide.

Experiment 6: To understand the effect of user behavior, we set up five browsers: instance A and B, the control pair, are configured identically except for the cookie that is needed for tracking user behavior. For C we browse 3 out of the 15 websites in the query set until Google learns the set of long-interests associated with those websites (which we verified[3]). For D and E we additionally browse random websites and perform Google searches on 50 product-related keywords shortly before collecting each snapshot (but don't click any result for D); we verified[3] that Google learned short-term interests for the random websites visited. Finally, for E we additionally click on product results before collecting each snapshot.

Figure 4(c) plots the similarity score between A and the other instances. While at first glance it might appear that browsing behavior, recent searches, and recent product clicks all result in different sets of ads, the problem is that even for the control pair similarity is very low — indicative of high levels of noise. We compared the fraction of ads shown to A and D that contained one or more of the search query terms and found no difference. The same held for ads containing the interests associated with instance C, and product names or categories used for instance E. We therefore believe that Google does not currently use recent browsing, search, or click behavior in picking website ads, but due to the high noise cannot definitively make the case for it.

That said, the measurement methodology developed here will allow us to monitor the evolution of these systems. As to the origin of the noise, we speculate this is because contextual ad systems have not yet been optimized for relevancy to the same extent as search ad systems, especially considering Google started collecting this data only since 2009 [9]. As contextual ad targeting improves over time, we expect the similarity score of the control pair to increase, and depending on whether the score for various user behaviors stay the same or increase in the future, we hope to conclude with certainty whether or not they are used.

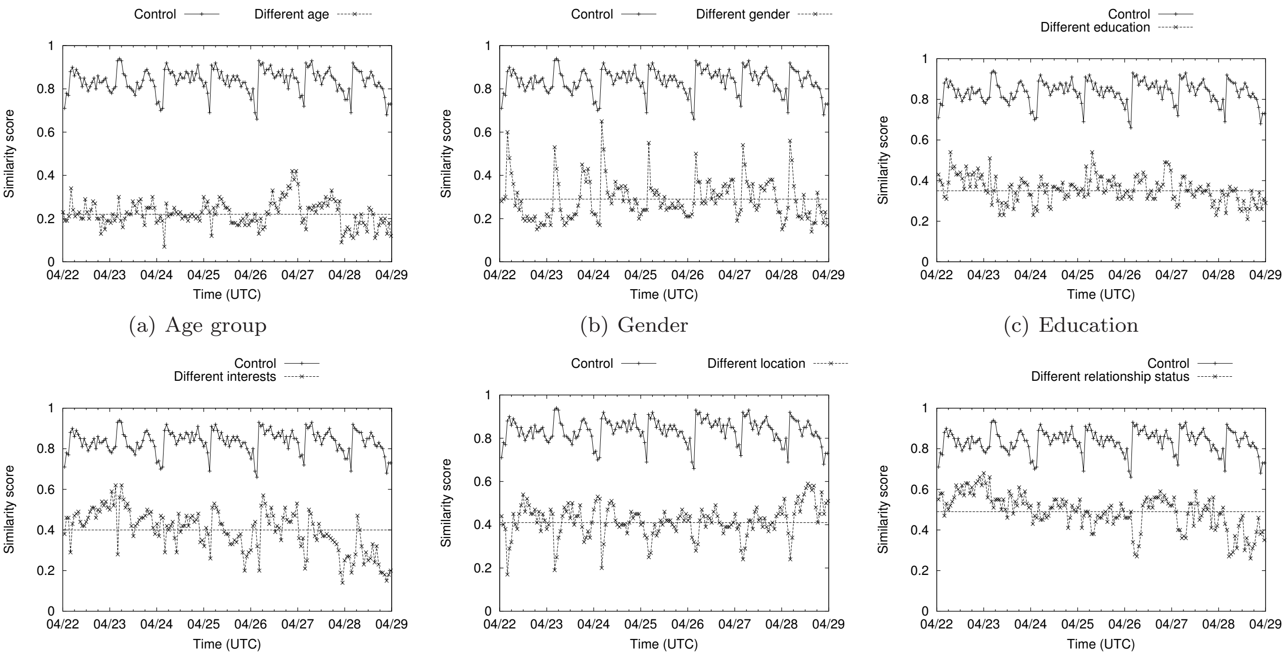
3.3 Online Social Network Ads

We next turn our attention to ads on online social networking sites, specifically Facebook. We seek to understand which pieces of profile information (gender, age, education, sexual-preference, etc.) Facebook uses today.

Experiment 7: We set up three (or more) Facebook profiles: profile A and B are set up identically (control), while profile C onwards differ from A in the value of the profile parameter of interest; the number of unique profiles depends on the number of values that parameter can take (e.g. 2 for gender, 5 for education, etc.) When not being varied, the gender was set to female, the age was set to 30, the location was set to New York, and the remaining fields were left empty.

Figure 5 plots the time-series of similarity scores for six different profile parameters. In short, Facebook uses all profile elements we checked. All the plots show some sort of diurnal behavior where around midnight US east-coast time the similarity score changes abruptly. Similar diurnal artifacts were observed in [5], and as in that paper, we believe the cause is the daily reactivation of ads that exhausted their daily budget the previous day. On the issue of profile elements, it appears that the two primary factors affecting ads are the user's age and their gender. While there exist values for education and relationship status that affect ads shown, not all values affect ads equally. For instance, there is little difference between ads targeted to users without any listed education and users in high-school. Similarly, if the gender is male, or the relationship-status is married, relationship-status has only a small impact on ads; the greatest impact of relationship-status on ads is seen for women who are engaged.

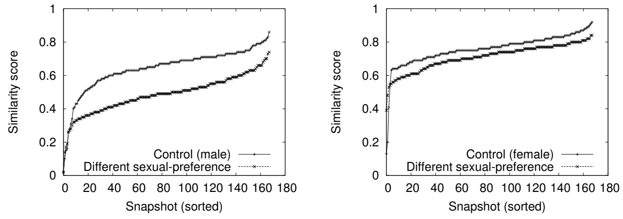
Experiment 8: Lastly, we set up six Facebook profiles to check the impact of sexual-preference: a highly-sensitive personal attribute. Two profiles (male control) are for males interested in females, two (female control) for females inter-



(a) Age group (b) Gender (c) Education

(d) Interests (e) Location (f) Relationship Status

Figure 5: Age group, gender, education, interests, location and relationship-status all appear to affect ads on Facebook



(a) Male (b) Female

Figure 6: Sexual-preference affects ads on Facebook, but more so for males than females.

ested in males, and one test profile of a male interested in males and one of a female interested in females. The age and location were set to 25 and Washington D.C. respectively.

Figure 6 plots the similarity scores for 1 week of data. While there is more noise in general, unlike in 4(c) there is a measurable difference between the control and test pairs; we further manually verified based on ad content that this difference is qualitative in nature (e.g. ads for gay bars were never shown for the control profiles, but shown often for the test profiles). The median similarity score for gay women was 0.15 higher than for gay men, indicating that advertisers target more strongly to the latter demographic.

Alarmingly, we found ads where the ad text was completely neutral to sexual-preference (e.g. for a nursing degree in a medical college in Florida) that was targeted *exclusively* to gay men. The danger with such ads, unlike the gay bar ad where the target demographic is blatantly obvious, is that the user reading the ad text would have *no idea* that by clicking it he would reveal to the advertiser both his sexual-preference and a unique identifier (cookie, IP address, or email address if he signs up on the advertiser's site). Furthermore, such deceptive ads are not uncommon; indeed exactly half of the 66 ads shown exclusively to gay men (more than 50 times) during our experiment did not mention "gay" anywhere in the ad text.

Overall we find that while location affects Google ads, behavioral targeting does not today appear to significantly affect either search or website ads on Google. Location, user demographics and interests, and sexual-preference all affect Facebook ads. As these systems evolve, our methodology can track changes in their use of user data. We thus inform, and hope to keep informed, the ongoing public debate about user privacy.

4. RELATED WORK

There is little past work in studying ad networks through measurement. In [5] we presented an ad hoc measurement result for Google ads, however, based on our experience presented here, we now believe that result significantly underestimated the number of ads and didn't properly account for noise.

5. SUMMARY

We have presented the first principled and robust methodology for measurement-based studies of online ad networks. We also inform the ongoing privacy debate regarding what user data is used today for targeting search ads, contextual ads, and ads on online social networks. Like most measurement studies, however, this analysis is a snapshot in time. Moving forwards, we hope that the methodology we have developed can continue to be used to broaden our knowledge of online advertising as well as to track trends in the future.

6. REFERENCES

[1] CoDeeN: A Content Distribution Network for PlanetLab. http://codeen.cs.princeton.edu/.

[2] M. Bolin, M. Webber, P. Rha, T. Wilson, and R. C. Miller. Automation and Customization of Rendered Web Pages. In *Proceedings of The Eighteenth Annual ACM Symposium on User Interface Software and Technology (UIST '05)*, Seattle, WA, Oct. 2005.

[3] DoubleClick. DART for Advertisers. http://www.doubleclick.com/products/dfa/index.aspx, 2009.

[4] J. Feng, H. K. Bhargava, and D. M. Pennock. Implementing Sponsored Search in Web Search Engines: Computational Evaluation of Alternative Mechanisms. *INFORMS Journal on Computing*, 19(1):137–148, Jan. 2007.

[5] S. Guha, A. Reznichenko, K. Tang, H. Haddadi, and P. Francis. Serving Ads from `localhost` for Performance, Privacy, and Profit. In *Proceedings of the 8th Workshop on Hot Topics in Networks (HotNets '09)*, New York, NY, Oct. 2009.

[6] B. Krishnamurthy and C. Wills. On the Leakage of Personally Identifiable Information Via Online Social Networks. In *Proceedings of The Seconds ACM SIGCOMM Workshop on Online Social Networks (WOSN '09)*, Barcelona, Spain, Aug. 2009.

[7] B. Krishnamurthy and C. Wills. Privacy Leakage in Mobile Online Social Networks. In *Proceedings of The Third Workshop on Online Social Networks (WOSN '10)*, Boston, MA, June 2010.

[8] B. Krishnamurthy and C. E. Wills. Cat and Mouse: Content Delivery Tradeoffs in Web Access. In *Proceedings of the 15th international conference on World Wide Web (WWW '06)*, Edinburgh, Scotland, 2006.

[9] Kurt Opsahl. Google Begins Behavioral Targeting Ad Program. http://www.eff.org/deeplinks/2009/03/google-begins-behavioral-targeting-ad-program, Mar. 2009.