What is a Cookie Worth?

Arslan Aziz and Rahul Telang

Heinz College, Carnegie Mellon University

June, 2015

Preliminary Draft

Abstract

Tracking a user's online browsing behavior to target her with relevant ads has become pervasive. There is an ongoing debate about the value of such tracking and the associated loss of privacy experienced by users. We inform this debate by quantifying the value of using different kinds of potentially privacy-intrusive information in targeted advertising. We find that using increasingly privacy-intrusive information increases the accuracy of prediction of purchases, but at a decreasing rate. We also find that targeted advertising is effective in increasing purchase probability and that this effect increases with the baseline purchase probability of a user. Finally, we simulate different privacy policy regimes by restricting different kinds of user information from being used for targeted advertising and quantify the impact such restrictions have on sales. We find that using privacy-intrusive user information can increase ad effectiveness by over 30% compared to random targeting. Using temporal information such as time spent by a user on an advertiser's website and time period since last visit increase ad effectiveness by over 20%. Other privacy-intrusive information, such as time spent on different types of product pages or number of unique products searched do not increase ad effectiveness significantly.

1. Introduction

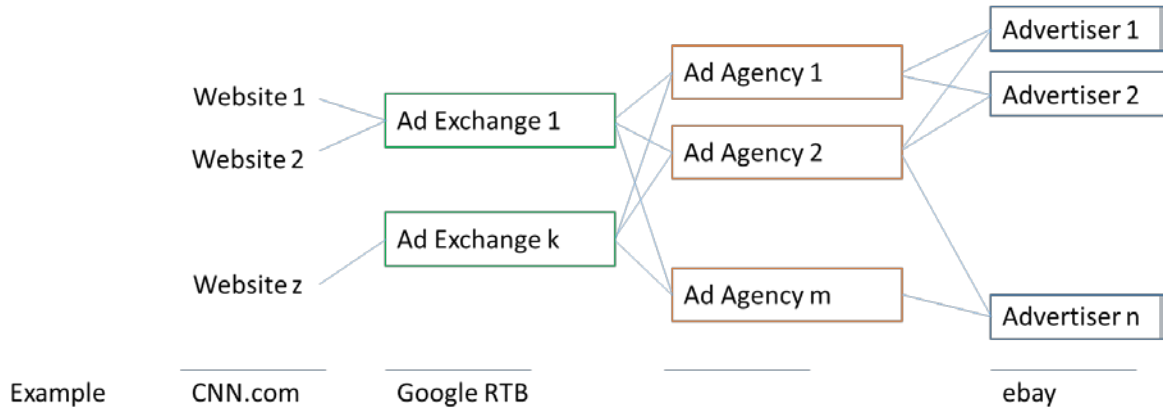
The online retargeted advertising industry has been growing rapidly in recent years. It is fueled by technological advancements that allow precise targeting of users with relevant and customized ads by tracking their online browsing behavior. This has led to concerns about the possible intrusion of privacy of users by such tracking and targeting technology.

In the absence of any policy directive, advertisers try to use as much information about an individual's browsing history as is technologically feasible. The argument on the advertiser's side is that increasing the relevance of an ad, by using sophisticated algorithms that targets users by only showing them products they are interested in, from brands they trust, and when they are receptive to them, benefit both the consumers as well as the advertiser. On the flip side, there are legitimate concerns that overt targeting comes at a significant loss of consumer privacy. Examples abound where firms target users causing significant privacy violation (at least as perceived by the consumers).[1] Several recent surveys have found that consumers are almost unanimous in their aversion to highly privacy invasive forms of advertising (Turow et al., 2009; Morales, 2010; Hoofnagle et al., 2012). In a recent 2014 Pew Survey on Public Perceptions of Privacy, 91% of U.S. adults say consumers have lost control over how their information is collected and used by companies and 64% believe the government should do more to regulate what advertisers can do with their personal information. Similarly, another survey by Harris Interactive, the TRUSTe Privacy Index finds that 92% of consumers worry about their privacy online, and 89% said they would avoid doing business with companies that do not adequately protect their privacy.

[1] In a widely noted and dramatic example, Target applied its data mining algorithm to determine whether women shoppers were likely pregnant. The store sent coupons for baby products to a pregnant high school student whose family did not yet know she was pregnant. “How Companies Learn Your Secrets,” The New York Times, last modified February 19, 2012, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all.

Online retargeted advertising, which refers to the delivery of personalized ads to users based on their online browsing history, is largely outsourced to specialized third party advertising firms that provide their clients user tracking and ad delivery service. In this paper, we refer to the firm whose products and services are being advertised as an 'advertiser' and the specialized third party advertising firm that provides the tracking and targeted ads as an 'ad agency'. The market has multiple advertisers that contract out advertising campaigns to different ad agencies. Multiple ad agencies compete in a real-time auction to win bids to show an ad to a user based on that particular user's browsing and online purchase history. These real-time auctions occur on exchanges that buy an inventory of advertising space on content websites, such as news sites or social networks, and in turn sell the advertising space to the highest bidder for each individual user. Figure 1 describes the retargeted advertising market structure. Websites sell ad inventory to ad exchanges, who sell the individual ad slots for a user to the ad agency who makes the highest bid. The ad agency bids on behalf of the advertiser for whose ads the user is likely to be most receptive.

Figure 1: Retargeted Advertising Market Structure



One of the key technologies that facilitates behavioral targeting is cookies. A cookie is a small piece of code embedded in a website that places a unique identifier in a user's browser when she visits a particular advertiser's website. This cookie enables the ad agency contracted by the advertiser to track the user's browsing behavior on that advertiser's website: what they view, what they click, what they purchase etc. When this user leaves the advertiser's website and browses some other content publisher's website, say a news website, the relevant ad exchange sends out a bid request to all available ad agencies, along with the unique identifiers in the cookies present in that user's browser. For example, suppose a user browsed an advertiser's website and then visited a news website. The news website would have sold the advertising space on its page to an ad exchange, and while the news page loads for the user, the ad exchange sends out a bid request to all the ad agencies in its database, along with the respective ad agency cookie identifier of the user for that advertiser. With this identifier, each ad agency is able to retrieve all the stored information of that user's online browsing behavior in the recent past and calculate the expected value of showing an ad to that user at that particular time and on the particular page. Note that the ad agencies are prohibited from using information from one advertiser while calculating the appropriate bid for another advertiser. So the only information an ad agency can use to make a bid to show an ad of a particular advertiser, is the information about the user's behavior on only that particular advertiser's website. It can also use information about the web page on which the ad is going to be displayed, a practice referred to as contextual targeting. The ad agency then sends this bid value to the ad exchange and the highest bidder wins the auction and gets to display its ad. Figure 2 describes the online journey that results in some ads being shown and others not.

Figure 2: Retargeted Advertising in Practice

Once an ad agency wins the bid to show an ad, it creates and serves an ad for that user based on the user's browsing history. This usually results in showing the user a mixture of products they have viewed in the past as well as other related products they might be interested in. This entire process, from the instant a user loads the news page, to the personalized ad being displayed by the ad exchange, takes only a few milliseconds. Appendix I shows some examples of personalized targeted ads that are displayed. It is clear that the user's browsing history shapes the ads that are ultimately delivered. This can lead to a perceived privacy intrusion in the consumer's mind, especially if the products displayed are sensitive products, such as medical products. The ads are also visible to other users of the same browser which is also a potential source of worry about loss of privacy. We note that tracking is done only on the particular advertiser's website and not on all the external websites that the user might browse. Ad agencies are contractually prohibited from using an individual's browsing history on one advertiser's website to target ads from another advertiser.

Policy makers worry about customer information being exploited for commercial gains by firms and agents unknown to them. This has led to a tug of war where policy makers would like to restrict (or in some cases stop) the use of information for explicit behavioral targeting while advertisers claim that such targeting generates large benefits for both the consumers and themselves. The current policy discussion surrounding “do not track law” in the US or “no cookie law” in Europe is an outcome of these concerns.

To make any policy recommendation, we need to quantify what is the value of different types of data collected via the cookie. We can then make a decision on whether to limit the use of this information or whether the benefits outweigh the costs. In this paper, we seek to quantify the economic benefit of tracking an individual via cookies by answering the following questions:

(i) How much do different types of information tracked via a cookie help in inferring a user’s purchase decisions?

(ii) What is the effect of advertising on the purchase decision and how does the baseline purchase probability moderate this effect?

(iii) What would be the impact of restricting certain types of information from being tracked via cookies on an advertiser’s overall sales?

We answer these questions in three steps: First, we note that if a user’s browsing information is of any value, it must be able to predict a user’s purchase intent. Greater use of potentially privacy-intrusive information is likely to increase the accuracy of predicting their purchases. We measure how the accuracy of predicting a user’s purchase probability increases when we include potentially increasingly privacy-intrusive information. We find that gains in predictive accuracy become harder as increasingly more information is used. In other words, after initial rapid gains in predictive accuracy by adding user information, adding further privacy-intrusive information to

predict purchases does not improve the prediction accuracy by much, and it may come at a privacy cost as perceived by the user.

Second, we quantify the effect of advertising on actual purchases made by individuals. Anytime a potential customer visits a website capable of displaying ads, an ad exchange requests a bid from all the advertisers on the exchange. Advertisers interested in showing an ad to that individual place their bid with the ad exchange, and the highest bidder wins and gets to show an ad. We estimate the effect of advertising by using a large dataset that contains details of every bid request that was made to all the potential customers of the advertiser under consideration. We have over 30 million bid requests for 586,909 unique individuals who are potential customers of the advertiser. We also have the amount that was bid for every bid request for each individual and whether the bid was successful or not. We also have information on whether these users made any purchase in the next three days. We utilize the bid amount for each individual as a measure of their baseline purchase probability as estimated by the advertiser. We use the number of bid requests received for a particular individual as a measure of the extent of their online activity during the period under consideration. This metric eliminates the 'activity bias' as described in Lewis, Rao and Reiley (2011). Previous observational studies that have tried to estimate the effect of ad have all been unable to capture the level of activity of an individual and have been plagued with severe biases and significant overestimation of the effect of ad. By controlling for activity, we eliminate this bias. In short, as researchers we observe all the information that the advertiser observes allowing us to estimate the effect of ads on purchase. In particular, we are interested in estimating whether the effectiveness of ads increases with higher willingness to purchase (as observed from the amount bid by the advertiser, which in turn is a function of information contained in cookies).

We find that cookie information predicts users purchase intentions effectively. We also find that effectiveness of an advertisement increases with the predicted purchase probability of a user. In short, users who are more likely to purchase (as inferred from detailed cookie information) are also most likely to be influenced by ads (though the magnitudes are small). In short, cookie based targeting is somewhat effective.

Finally, we conduct a counterfactual simulation where we remove some cookie information (say more privacy invasive information) and compare the targeting effectiveness with when all cookie information is present. We find reducing the extent of information available for tracking decreases the effectiveness of online advertising and allows us to estimate the tradeoff between ad effectiveness versus consumer privacy.

2. Literature Review

The downside of restricting the use of cookie information is that it will adversely affect the multi-billion dollar ad industry and slow down its innovations. The effectiveness of traditional untargeted forms of advertising such as television ads, billboards or internet banner ads has proven to be very hard to measure (Lewis & Rao 2014). Targeted advertising promises to reduce wasted advertising spend by letting advertisers pick and choose which ad to show, when to show it and whom to show it to. Unfortunately, empirical work in demonstrating the value derived from cookie based targeting is sparse at best. Golfarb and Tucker (2011) show that in Europe, after passage of the 'no cookie law' restricting the use of personal information for targeting, the effectiveness of ads reduced. However, they did not have any measure of (i) what cookie information is used by advertisers for targeting purposes, (ii) actual click and purchase data. They could only measure

purchase intentions. Lambrecht and Tucker (2013) find that generic retargeted ads, where the decision to show an ad or not is based on users' browsing history, but the content of the ad is not tailored are on average more effective than ads that are both targeted and personalized. They further demonstrate that this effect is dependent on which stage of the purchase decision cycle a consumer is in. For consumers closer to making the purchase decision, a personalized ad is more effective than a generic one, and vice-versa. They use a visit to a travel review site as a measure of how close a consumer is to making a purchase decision. The dependent variable in this study is click through rates of advertisements. Visiting a product review website would be a binary and noisy measure of how close a consumer is to making a purchase and click through rates in online advertising have been shown to be poor predictors of actual purchases by consumers. As we will show, we overcome both these challenges.

Budak et al. (2014) suggest that 'Do-not-track' legislation would impact content providers but suggest that the shortfall in revenues due to decreased effectiveness of advertising could be made up by switching to a "freemium" model. Johnson (2013) estimates the financial impact of different privacy regulations on online publisher and advertiser revenues. Both these papers compare the impact of policies that either entirely restrict cookie based tracking, allow it fully or allow users the choice to do so.

Most research and policy discussions to date have assumed a binary decision of whether cookie based tracking should be made permissible or not, and has ignored the variety of information that a cookie can track. This has led to policies such as the 'EU no cookie law' that restrict the use of any cookies based information or require the explicit consent from a user before using them. However, it has been recognized that cookies are also useful to deliver basic website functionalities such as keeping a user signed in to a website. The different types of information that can be tracked

via a cookie have different values for advertisers as well as different perceived privacy costs. An effective privacy policy must weigh the benefits of tracking different types of information against the associated privacy costs and make a decision based on this tradeoff. If the privacy cost of using some piece of information tracked via a cookie is low, and the associated benefits from improved relevancy of ads is high, a sophisticated privacy policy should allow such tracking. If, on the other hand, some pieces of information tracked via a cookie are perceived as very privacy intrusive, while the associated benefit from using this information to target ads is low, then such information tracking should be restricted.

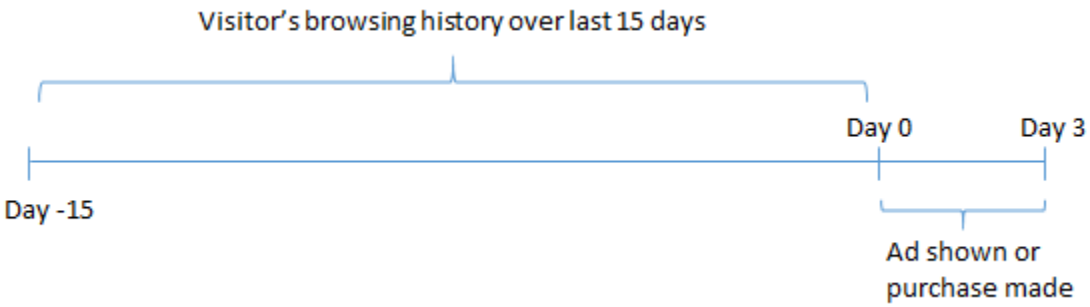
However, there is almost no research that connects information in user cookies to their purchases and how an advertiser uses this information to bid for an ad. This is the critical gap in the literature that we try to fill with this paper.

3. Data

We have collaborated with a large digital advertising firm with over 200 clients spread over 20 countries. We were granted exclusive access to proprietary data for one large advertiser. For this advertiser, we collected multiple datasets.

In the first dataset, we have data on all the visitors of the mobile website of a large e-commerce firm, specializing in apparel, in the period between April 20 and May 5, 2014. There were 248,808 unique visitors to the site during this 15 day period. Our data consists of very detailed website browsing and transaction history of these visitors over the past 15 days, as well as the purchases made by these visitors in the next three days. These users made 1,567 purchases from May 6 to May 9th, 2014. This is shown in Figure 4.

Figure 4: Timeline of data capture



The browsing and purchasing information are all tracked via a cookie stored on a visitor's browser by the third party ad agency when a user visits the advertiser's website for the first time. From then on, all the subsequent user interactions are tracked and added to the information being stored for that particular user. Cookies are a common way to identify the uniqueness of a user, though the same user can visit or transact at the website through multiple devices. Our data, though, is limited to the mobile website of an e-commerce firm. We note that cookie ids are not mapped to any personally identifiable information and so our study does not infringe on any individual's privacy. All we observe is that some individual browses certain products and makes certain purchases, but there is no way for us, or the advertiser, to know who that particular individual is.

Users can navigate on the website via – (i) homepage, (ii) category page (such as Men, Women, Sportswear, Formal wear etc.) (iii) product pages (individual products are shown), and (iv) shopping cart. We are able to track behavior on each of these types of pages separately. We group all the variables describing a user's browsing history into six levels of information.

1. Non-behavioral variables: These variables do not track the browsing history of the user and contain only variables regarding the manufacturer, browser and operating system of the device being used to visit the website. This represents the case when no browsing behavior

is available for tracking and targeting. Note that this doesn't include any personally identifiable information like IP address, location, email id, phone number or any other information from which an individual may be identified by any means.

2. Website level: These variables include information about whether a user has visited a website in a given time period, whether he has made a purchase or added a product in the shopping cart, or if she has logged in or performed a search on the site.
3. Product level: These variables include information regarding the behavior of the user within the website during each session in the last 15 days, such as the number of product or category pages visited, number of products added to the shopping cart or purchased etc.
4. Temporal website level: These variables capture the total time spent on the entire website and the time since last visit for a user.
5. Temporal product level: These variables capture the time spent on each type of page (homepage, category pages, product pages etc.).
6. Product and category frequency: These variables capture the number of times the most often visited and most recently visited product categories were visited by the user.

We describe the variables in Table 1 and Table 2 provides the summary statistics.

Table 1: Description of variables

Type	**Variable**	**Description**
Non-browsing variables	Cookie	Unique tag for each mobile website visitor
	Device	Mobile device used by visitor
	Browser	Browser used by visitor
	OS	Operating System of the device
Website level	n_sessions	Number of unique sessions by a visitor in last 15 days.
	n_views	Number of visits by a visitor in last 15 days.

	Cart_flag	Whether any product was put in shopping cart
	Purchase_flag	Whether any product was purchased
	Search_flag	Whether user performed a search on the website
	Login_flag	Whether the user logged in to the website with their account
	Click_flag	Whether the user clicked on an ad to reach the website
Product level	Homepage_views	Count of number of visits to the advertiser's homepage
	Category_views	Number of category page visits
	Product_views	Number of product page visits
	ShopCart_views	Number of shopping cart visits
	Uniq_cat	Number of unique categories visited
	Uniq_subcat	Number of unique sub-categories visited
	Uniq_prod	Number of unique products visited
	Cart_count	Count of number of products added to shopping cart
	Search_count	Count of searches performed by user on the website
	Click_count	Ads clicked by in the last 15 days
	Purchase_count	Count of products purchased by user in last 15 days
Temporal Website	Tos	Total time spent on site in last 15 days (in minutes)
	Hours_dropoff	Hours since last visit to the advertiser's website
Temporal Product	Homepage_tos	Time spent on homepage (in seconds)
	Category_tos	Time spent on category pages (in seconds)
	Product_tos	Time spent on product pages (in seconds)
	ShoppingCart_tos	Time spent on shopping cart (in seconds)
Most seen and last seen frequency	Most seen frequency	Number of times the most seen category, subcategory and products were seen in the browsing period
	Last seen frequency	Number of times the last seen category, subcategory and products were seen in the browsing period
Dependent Variable	Purchase_next3	Purchase indicator for the next 3 days.

Table 2: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
logged_in	248,808	0.017	0.130	0	1
n_views	248,808	52.606	107.404	2	6,993
tos	248,808	43.518	84.829	0.000	3,119.820
n_sessions	248,808	1.000	0.009	1	2
homepage_views	248,808	3.775	7.427	0	444
category_views	248,808	17.755	42.670	0	1,919
subcategory_views	248,808	11.210	22.754	2	1,818
product_views	248,808	2.335	8.804	0	1,038
homepage_tos	248,808	198.024	675.536	0	36,046
category_tos	248,808	647.731	1,717.398	0	86,268
subcategory_tos	248,808	681.514	1,898.586	0	102,680
product_tos	248,808	206.289	971.455	0	60,152
uniq_cat	248,808	0.272	0.704	0	10
uniq_subcat	248,808	3.914	2.786	0	50
uniq_prod	248,808	8.234	15.107	0	820
last_seen_cat_freq	248,808	1.618	9.004	0	864
last_seen_subcat_freq	248,808	16.812	36.852	0	1,862
last_seen_prod_feq	248,808	1.614	1.547	0	159
most_seen_cat_freq	248,808	1.878	10.021	0	864
most_seen_subcat_freq	248,808	21.729	44.122	0	3,153
most_seen_prod_freq	248,808	2.306	2.415	0	189
click_flag	248,808	0.250	0.433	0	1
search_flag	248,808	0.081	0.272	0	1
cart_flag	248,808	0.296	0.456	0	1
has_bought	248,808	0.099	0.603	0	111
hours_dropoff	248,808	133.478	101.841	0	360
purchase_flag	248,808	0.006	0.079	0	1
cart_count	248,808	0.463	0.904	0	4
click_count	248,808	0.288	0.542	0	7
search_count	248,808	0.130	0.584	0	25
ads shown in browsing period	248,808	6.536	17.538	0	889
ads shown in purchase period	248,808	1.465	6.506	0	450

The second dataset is a bid request level data for all the potential customers of the advertiser under consideration over a period of three consecutive days. We have over 30 million bid requests, each with a bid amount, for 586,909 unique individuals. These individuals saw a total of 961,166 ads during the three day period and made 3,866 purchases from the advertiser. We aggregate this

dataset at a unique individual level and use the average of the bid amounts for that individual as a measure of her estimated baseline purchase probability and the number of bid requests received as a measure of the user's online activity level. We can then estimate the effect of the ad and its interaction with the baseline purchase probability after controlling for both baseline purchase probability and online activity.

4. Analysis

Our analysis proceeds in three steps. First, we quantify the improvement in predictive accuracy when more information is used to predict the purchase probability of an individual. Next, we quantify the effect of seeing an ad on the individual's purchase probability and how this effect is moderated by the baseline purchase probability of an individual. Lastly, we estimate the impact of counterfactual privacy policies that restrict certain kinds of information from being utilized for targeted advertising by quantifying the loss in potential sales such policies cause by lowering advertising effectiveness.

4.1 Predicting Purchases from Online Browsing Behavior

We use logistic regression models to predict purchases based on observed browsing behavior on the website over the last 15 days by each user. We use six model specifications that make use of increasingly privacy intrusive trackers of user's browsing and transaction behavior to predict future purchases.[2]

[2] We include analysis of a different model specifications in Appendix II

In the first model, we use purely non-behavioral information to predict the probability of purchase for an individual. This is the 'no privacy intrusion' scenario which we use as a base case to compare the improved accuracy of using cookie information in predicting probabilities of purchase.

In the second model, in addition to the non-behavioral variables used in the first model, we use aggregate website-level metrics such as number of visits, whether products were added to the shopping cart, whether the user searched on the website, made a purchase, or logged in to the website.

The third model, in addition to the variables used in the second model, uses information regarding the browsing behavior of the user within the website, such as number of homepage views, number of category page views, number of product page views etc. In addition, we use counts instead of indicators for variables such as products added in the shopping cart, searches performed or products bought. We believe these are more privacy intrusive as information regarding the specific pages visited by the user is tracked and utilized by the advertiser.

The fourth model, in addition to variables used in the third model, uses temporal data regarding time spent by the user on the website, such as total time spent on site and time since last visit.

The fifth model, in addition to variables used in the fourth model, uses temporal data according to each webpage type, such as the amount of time spent on category pages or product pages.

The sixth model, in addition to variables used in the fifth model, uses information regarding the number of times the most seen and last seen product or product category has been visited by the user in the last 15 days. This is the most privacy intrusive data that the advertiser can use given the data they track using cookies.

We note that since in each model, we use all the variables in the previous model and add a few additional variables, each subsequent model is more privacy intrusive than the previous one.

We use a randomly drawn sample of 75% of our total sample as our training dataset and train logistic regression models on the data. We then use the remaining 25% of the sample as the test dataset on which we predict the accuracy of the different predictive models. We then compare the accuracy of these different predictive models using the Area Under the receiver operating characteristic Curve (AUC). This is summarized in Table 3.

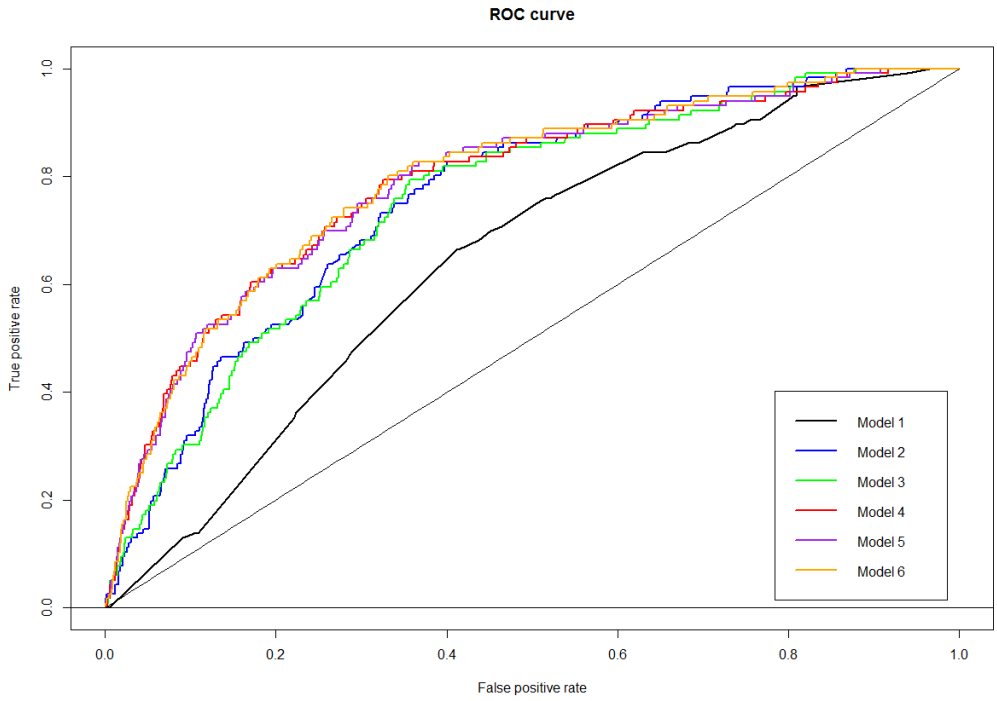
Table 3: Model Comparison

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	Non-behavioral variables	Model 1 + Website level variables	Model 2 + Product level variables	Model 3 + Website Temporal variables	Model 4 + Product Temporal variables	Model 5 + Last seen/Most seen frequency variables
AUC	0.5654682	0.7529817	0.7576667	0.8134579	0.81296	0.814859

The AUC is an appropriate metric to compare the relative accuracy of prediction of different models as it captures the ability of a model to discriminate between purchases and non-purchases irrespective of the threshold used to classify a predicted probability of purchase as a purchase or non-purchase. We find that the AUC increases with increasing use of privacy intrusive variables but at a decreasing rate. In particular, we find that the AUCs for models 2 and 3 are very similar, as are the AUCs for models 4, 5 and 6. This suggests that the usage of the additional product level variables in model 3 and the product temporal and frequency variables in models 5 and 6 do not provide any significant increase in predictive accuracy. So, from a privacy policy perspective, the set of models that need to be considered are only Model 1, Model 2 and Model 4, since Model 3 has a similar predictive accuracy as Model 2, but uses more privacy-intrusive information, and

Model 5 and 6 have similar predictive accuracies as Model 4 but with additional privacy-intrusion. We plot the ROC curves for all the six models in Figure 5.

Figure 5: ROC Curves for different models



The diminishing return in predictive accuracy to the usage of increasing privacy intrusive variables for targeting suggests that even when firms have access to highly intrusive information about an individual's behavior, it might not lead to a significant improvement in predictive ability. There is a possibility that a privacy policy that prevents firms from collecting and utilizing such information can do so without harming the effectiveness of targeted advertising. In fact, making users aware of the fact that some of their information is intentionally not being tracked might not only improve the effectiveness of advertising, it might reduce their privacy concerns and may make them more receptive to online advertising. This result has been suggested in Tucker (2014), where they show that when Facebook announced improved privacy controls for users, the effectiveness of targeted advertising on Facebook increased.

4.2 Effect of Advertising on Purchases

In this section we estimate the effect an ad has on an individual's purchase probability and whether this effect is influenced by their inherent baseline purchase probability. If we find that ads have an effect, and that this effect doesn't vary with the baseline purchase probability of an individual, then targeted advertising would not be socially useful. Advertisers still may be interested in targeting users but targeting generates no social good (but imposes costs due to privacy violations). If, instead, we find that the effect of ad is moderated by an individual's baseline purchase probability, then it is important for the advertiser to able to identify and target the individuals for whom the ads are most effective.

Extant literature has pointed out the difficulties of measuring the effect of advertising. Lewis, Rao and Reiley (2011) point out the difficulty in measuring the effect of advertising through observational data. The main issue is that the extent of user's online activity is an endogenous variable that is strongly correlated with both the probability of seeing an ad and of making an online purchase. The difference between the purchase rates of the samples of individuals who see ads and those that don't see it measure not just the effect of the ad, but also the difference in the online activity level of the two groups. Lewis and Rao (2014) point out that even large scale randomized control trials are not sufficient to estimate the effect of ads.

In this paper, we take a different approach to getting rid of the endogeneity caused by the online activity of the individual. By using an aggregate of the number of bid requests were received by the advertiser for each potential customer, we are able to measure the level of online activity of the individual. By controlling for bid request, we control for the extent of online activity of each individual and hence eliminate this potential source of unobserved variable bias. We should reiterate that we as econometricians, observe the same information that the firm does (i.e the firm

knows the value of a customer via its cookie and bids an amount proportional to it, it knows the activity level of the customer via the number of bids requests it receives from the ad exchange).

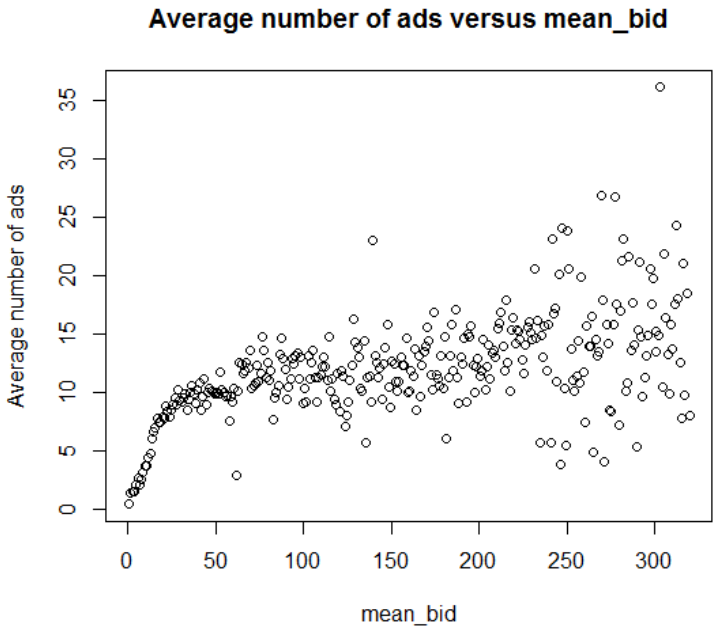
Since the firm may receive many bid requests for a customer, we take the average of these bids as a measure of the firm's valuation of a particular customer. We also know the number of ads the individual was shown and whether the individual made a purchase from the advertiser during the period under consideration. To create a subset of only potential customers, we drop the users who had an average bid value equal to the lowest default bid placed by the advertiser for every user. These users, who are almost 15% of the dataset, have not visited the advertiser's website over the last 30 days and hence the advertiser doesn't want to show an ad to them. This reduced dataset is referred to as dataset 2. Table 4 summarizes dataset 2.

Table 4: Summary of Bid level dataset

Number of unique individuals	498,036			
Number of ads shown	956,805			
Number of individuals who see at least one ad	103,765			
Number of individuals who click at least one ad	4787			
Number of ads clicked	8892			
Number of individuals who make at least one purchase	3465			
	Mean	St. Dev.	Min	Max
Mean_bid amount (USD per thousand ads)	12.24	41.36	0	320
Number of bid requests	54.61	475.52	1	227,693
Number of ads per individual	1.92	12.22	0	5,206
Number of ads per individual, given they see at least one ad	9.22	25.48	1	5,206

On average the firm bids about 12 USD per thousand ads, for each individual, though the bid values have a large variation. It also receives on average 55 bid requests for each active user with a very high variance as seen in the data. This represents the variation in online activity of individuals. The firm shows about 9.22 ads, on average, to an individual who sees at least one ad.

In Figure 5, we plot mean bids with the number of ads shown.



Not surprisingly, users with larger mean bids are more likely to see more ads but the effect tapers off quickly. Most users are likely to see 0-10 ads. Recall majority of users have small mean bid values.

In Figure 6, we show the proportion of users who make a purchase versus the average bid value placed for the user. We see that the average bid value is strongly correlated with increasing purchase probability. This suggests that firm correctly bids more for users who are more likely to purchase. [3]We also plot the histogram of number of bid requests received for those that receive less than 100 bid requests and note that it is an exponentially decreasing distribution.

[3] AS we noted earlier, cookie information is the only information firm has about the end user.

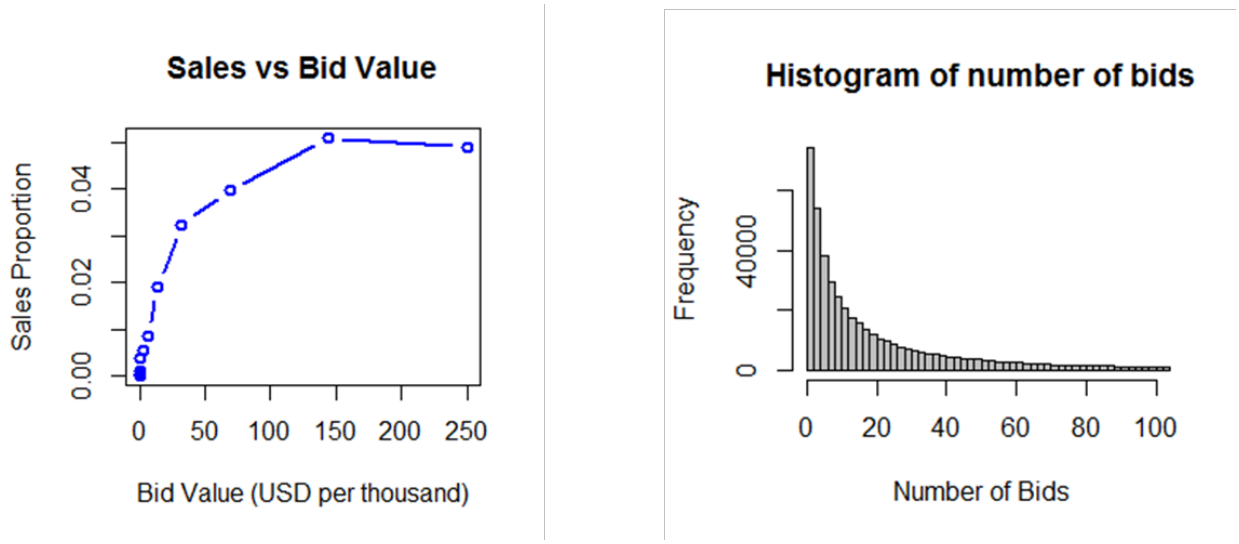


Figure 6: Sales vs Bid Value and Histogram of number of bids

To estimate the effect of ad, we estimate the following linear OLS regression:

$$pur_i = \beta_0 + \beta_1 * lmean\_bid_i + \beta_2 * n\_ads_i + \beta_3 * n\_bids_i + \beta_4 * lmean\_bid_i * n\_ads_i + \epsilon_i \quad (1)$$

i indexes a user (or cookie). *pur* represents the number of separate purchase transactions by an individual in the three day period under consideration and ranges from 0 to 12 in our sample. *Mean_bid* is the average amount bid by the firm for the user and captures the firm's valuation of the user. We use the logarithm of the *mean_bid*, denoted by *lmean_bid* because of the large variance of *mean_bid* values and the potentially non-linear effect the ad has for different values of *mean_bids*. A higher *mean_bid* implies an individual has a higher baseline purchase probability and should lead to higher purchase rate. *n_bids* captures how active the user has been on the Internet during the purchase period and is measured by the number of bid requests for the user received by the firm. *n_ads* is a count of the number of ads seen by an individual. The interaction of *lmean_bid* with *n_ads* captures the idea that whether the users who are more likely to purchase

are also more influenced by the ad. Finally, even though we can estimate Logit or Probit regressions, given the interaction effect β_4 we are interested in, we use a linear model.[4]

Identification

It is useful to go over identification assumptions one more time in Eq (1). The model assumes that the after controlling for mean_bids and number of bids, the number of ads that a user sees are random. Or, that there are no systematic unobervables (to the econometricians) that are correlated with the ads shown. Based on our discussion with the firm, it was made clear that the firm has no information about users other than what they have in the cookie. It does not account for the value of the goods when bidding for a user. It was also made clear that the website on which the user will be shown the ad does not play any role in its bidding algorithm[5]. Given that cookie information is a noisy measure of users' value, it is much more likely that the firm keeps varying its bids. Thus the mean_bid potentially captures the firms value of a user quite well and the respective bid around the mean are likely random. Sometimes it is successful in winning, while it fails in other attempts. In some cases, the firm needs certain campaign level targets to be fulfilled and bid or less depending on those requirements. Finally, sometimes it may lose the bid due to competitive reasons. So for the same mean bid value, one user may see more (or less) ads from the firm depending on competing firms' bidding strategies. In summary, we believe that conditional on observables, the number of ads a user sees is uncorrelated with the error term. We report the estimation results in Table 5.

Table 5: Results

[4] Interpretation of interaction terms in non-linear models is particularly challenging.

[5] We still might worry about it that a manager might make manual decisions to calibrate a bid. Even though the advertising firm has information about which website the ad will potentially be shown on, according to the firm and from an analysis of their bidding algorithms, this information is not utilized while bidding.

	Dependent variable:
	sale
log(mean_bid)	$2.321 \times 10^{-3***}$
	(1.458×10^{-4})
number of ads	$2.788 \times 10^{-4***}$
	(1.903×10^{-5})
number of bids	-2.475×10^{-7}
	(4.897×10^{-7})
log(mean_bid) x number of ads	$1.854 \times 10^{-4***}$
	(5.598×10^{-6})
Constant	$8.542 \times 10^{-3***}$
	(1.458×10^{-4})
Observations	498,036
R^2	0.0167
Adjusted R^2	0.0167
Residual Std. Error	0.09303 (df = 498031)
F Statistic	2,116*** (df = 4; 498031)
Note:	*p<0.1; **p<0.05; ***p<0.01

We note that except for the number of bids (*n_bids*), all of the coefficients are positive and statistically significant. Notice we have a very large sample giving us precision. Higher bids are placed for users who are more valuable (as estimated from cookie data) and are more likely to purchase. The effect of each individual ad is positive though very small. More importantly, the interaction term between *lmean_bid* and *n_ads* is positive and statistically significant. Each additional ad increases the purchase probability by $\beta_2 + \beta_4 * lmean\_bid$. For the average

individual, *lmean_bid* = -1.1, which implies that each incremental ad increases purchase probability for the average individual by 0.008%. Viewing a 100 ads increase the purchase probability of an individual by about 0.8%. Note that these are probabilities expressed as percentages. The effect of the ad is increasing with bid value, which is a measure of the baseline purchase probability of an individual. Hence, advertisers would wish to target individuals more likely to purchase their products, as these individuals are the ones who are most influenced by ads.

With these two estimates in hand, we can now do policy experiments on how restricting cookie information would affect ad effectiveness and advertisers' willingness to pay.

4.3 Advertiser Revenues versus Consumer Privacy

In this section, we calculate the value of a consumer's browsing information by calculating the loss in sales that would result from imperfect targeting were that information made unavailable to the advertiser. We have shown that having access to a lesser amount of consumer information would lead to less accurate predictions about a user's baseline purchase probability. In the previous section, we established that a person's response to an ad increases with her baseline purchase probability. In this section, we combine these two results and run counterfactual privacy policy simulations in which we restrict some of the consumer's browsing information and measure the loss in advertising effectiveness.

From the first dataset, we predicted a user's baseline purchase probability from different logistic regression models that made use of increasingly privacy-intrusive user information. We showed that the accuracy of these predictions increased with the use of increasing information. We now call the full-information predictive model (Model 6) as the 'true baseline purchase probability' of

the individual. It is the best prediction we have about a user's intent to purchase and denote this as *M6 pred*.

In the second dataset, we used the average bid placed for an individual as a measure of their baseline purchase probability. We used this to measure the effect of the ad after controlling for the individual's baseline purchase probability. We can normalize the *mean_bid* value by dividing each with the maximum *mean_bid* in the dataset. We call this *norm_bid*, and note that its value ranges from 0 to 1 and is a measure of an individual's baseline purchase probability.

Our estimation of the decrease in ad effectiveness from the use of less user information is as follows: We first find a mapping between the two measures of an individual's baseline purchase probability – *M6 pred* and *norm_bid*. We then estimate the effect of an ad on sales in terms of *norm_bid* using dataset 2. We then replace *norm_bid* with the corresponding *M6 pred* for each individual in dataset 1. We then assume a targeting strategy that involves the advertiser targeting individuals with the highest baseline purchase probability to maximize the effect of ads. We construct six counterfactual privacy policies that allow user information in each of the six predictive models respectively to be used by the advertiser. We note that the higher the accuracy of the predictive model, the more accurate the advertiser will be in identifying the highest baseline purchase probability, and the more effective the ads. We then estimate incremental sales for each predictive model by using the estimated baseline purchase probability *M6 pred* and compare. These differences give us the estimate of ad effectiveness versus consumer privacy.

Before we proceed, we define a new variable in dataset 2 called 'sale' which is a dummy variable with a value 1 for every individual who makes at least one purchase and 0 otherwise. We do this since we only have a 'sale' dummy variable in dataset 1 and not the number of purchase transactions. Also, we convert the *n_ads* variable in dataset 2 to a dummy variable *ad* with a value

of 1 if an individual sees at least one ad and 0 otherwise. This is done to simplify our calculations and avoid making assumptions about the number of ads individuals would see in our counterfactual scenarios.

To find the mapping between *M6 pred* and *norm_bid*, we plot the proportion of individuals served ads and proportion of individuals who make purchases as a function of the *M6 pred* and *norm_bid*. These are shown in Figure 7.

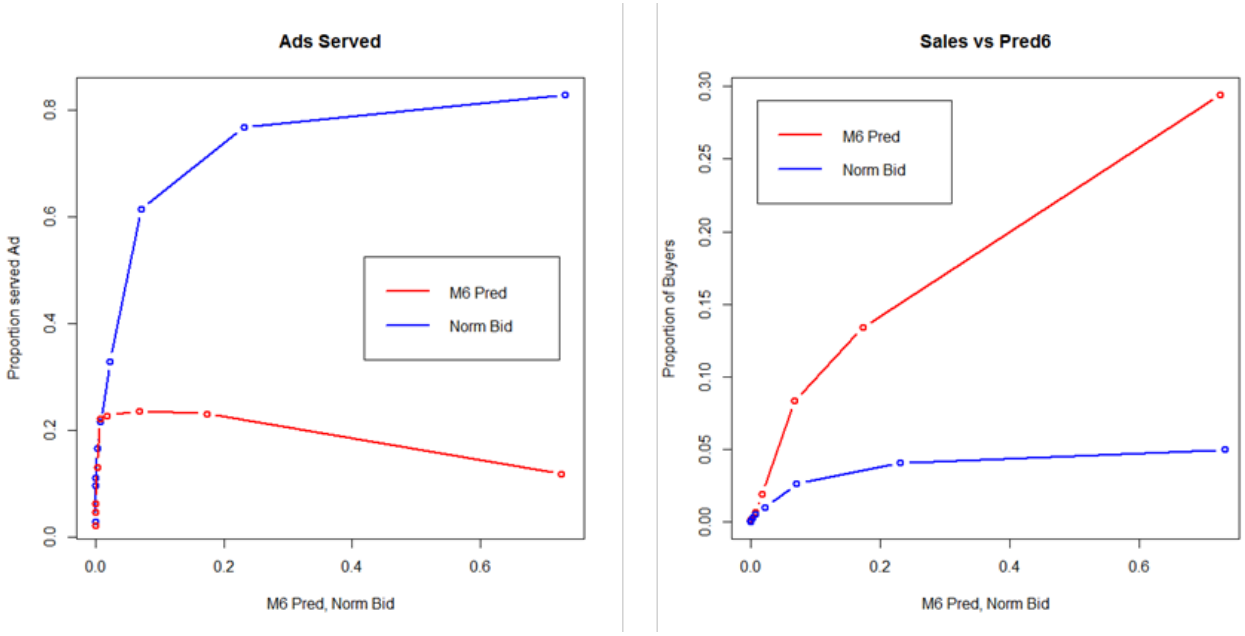


Figure 7: Ads served and Sales versus measures of baseline purchase probability

We find that over most of the range of *M6 pred* and *norm_bid*, the ads served and sales proportions vary significantly. In particular, we observe that the probability of being shown an ad is accurately predicted by the *norm_bid*, as would be expected, while the probability of making a purchase is better predicted by *M6 pred*, again, as expected. We also note that for higher values of *M6 pred*, the probability of being shown an ad decreases with increasing *M6 pred*. This suggests that the advertiser is missing out on showing ads to such individuals due to inaccurate targeting. But if we

focus on the just the lower end of the range, such that *M6 pred* and *norm_bid* lie between 0 and 0.01, we find the two estimates of baseline purchase probability are very similar. We plot this in Figure 8.

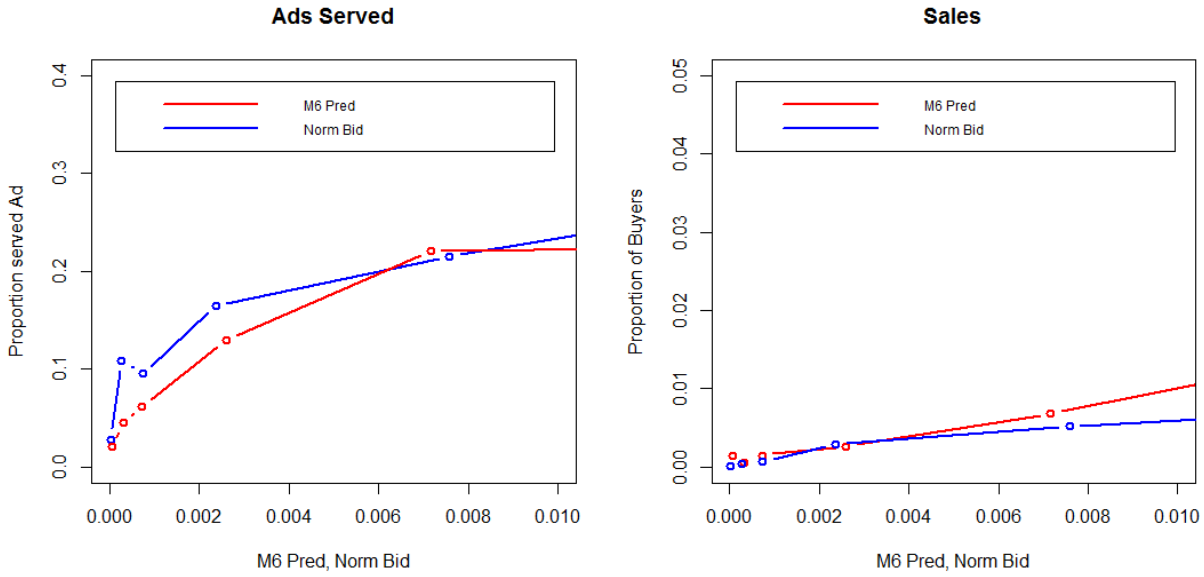


Figure 8: Ads served and Sales over measures of baseline purchase probability

Within this range, we can assume *M6 pred* and *norm_bid* are equal on average. We note that this range of values constitute the majority of individuals and a sizeable share of individuals who see ads and make purchases. For dataset 1, which we will use to estimate the ad effectiveness for different privacy policy regimes, this range includes 85% of all individuals, 75% of individuals who see at least one ad and almost 40% of individuals who make a purchase. We also note that the 6429 individuals who were shown ads comprise 12.16% of the total sample. These are summarized in Table 6.

Table 6: Characteristics of selected sampled (*M6 pred*, *norm_bid* < 0.01)

Pred, norm_bid < 0.01	N	% of Total	Ads	% of Total	Sales	% of Total
Dataset 1	52,856	85%	6429	75%	158	39.5%
Dataset 2	373,266	74.9%	40,257	38.8%	616	17.8%

We now find the effect of the ad on sales, using *norm_bid* as the control for the baseline purchase probability in dataset 2, with the following linear OLS regression:

$$sale_i = \gamma_0 + \gamma_1 * lnorm\_bid_i + \gamma_2 * Ad_i + \gamma_3 * nbids_i + \gamma_4 * lnorm\_bid_i * Ad_i + u_i$$

where *lnorm_bid* is the logarithm of the normalized mean bid. The results of this regression are shown in Table 7. Since we use a different dependent variable (sale dummy) and an ad dummy instead of number of ads, we cannot directly compare the coefficients with those in Table 5.

Table 7: Regression results

	Dependent variable:
	sale
log(norm_bid)	$2.823 \times 10^{-4***}$
	(2.745×10^{-5})
ad dummy	$2.563 \times 10^{-2***}$
	(8.692×10^{-4})
number of bids	1.723×10^{-7}
	(1.219×10^{-7})
log(norm_bid) x ad	$2.806 \times 10^{-3***}$
	(1.259×10^{-4})
Constant	$3.191 \times 10^{-3***}$
	(2.412×10^{-4})
Observations	373,266
R^2	0.005
Adjusted R^2	0.005
Residual Std. Error	0.040 (df = 373261)
F Statistic	504.675*** (df = 4; 373261)
Note:	*p<0.1; **p<0.05; ***p<0.01

As discussed earlier, we can replace *log(norm_bid)* with *pred* in the results in Table 7. We note that, as in Table 5, except *number of bids*, all the other coefficients are positive and statistically significant.

Our counterfactual privacy policy simulation is as follows: In dataset 1, 12.16% individuals see an ad. We now consider six scenarios in which advertisers have access to user information included in each of the six predictive models respectively as in Table 3. In each scenario, we assume the

advertiser uses only the information available to it to predict each individual's baseline purchase probability and then targets the top 12.16% individuals with the highest estimated baseline purchase probability and shows them ads.

Then we calculate the effect of an ad, as

$$\Delta Sales_i = E(Sales|Ad = 1)_i - E(Sales|Ad = 0)_i = \gamma_2 + \gamma_4 \times \log(pred_i)$$

where $pred_i$ as the baseline purchase probability rate given by the full information model (M6).

Aggregating over a sample of T individuals, we can calculate the incremental sales for that sample

$$\text{Incremental Sales} = \sum_{i=1}^{T}(\Delta Sales_i) = T * \gamma_2 + T * \gamma_4 * (avg(\log(pred)))$$

where *avg.pred* is the average baseline purchase probability of the sample.

To quantify the value of targeting under each of our six counterfactual privacy policies, we subtract the sales that would have resulted from random display of ads.

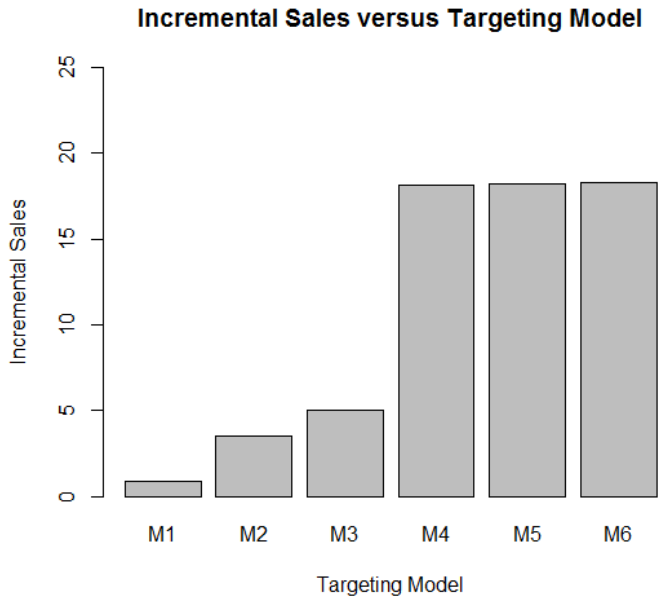


Figure 9: Incremental Sales versus Targeting Model

We find that targeting using only non-behavioral variables, in Model 1 (M1) leads to a small incremental sale. We see moderate increases in incremental sales in Model 2 and 3 and then a significant jump for Model 4. Models 5 and 6 have similar incremental sales. Note that random targeting of individuals with ads would have led to 60 incremental sales attributable to the ads. So, in terms of overall incremental sales due to targeted ads, we find that when the full information model (M6) is used to target users, ads are 30.5% more effective than random targeting of ads and 28.7% more effective than non-behavioral targeting. In consecutive models, the most significant jump in incremental sales of 20.3% occurs from M3 to M4, when two temporal website level variables are included – time spent on the website and hours since the last visit. Adding detailed temporal variables that break up the time spent by a user on the website by type of page visited and including estimates of the frequency of visiting a page and the breadth of browsing don't increase the impact of ads significantly.

Our results demonstrate that certain kinds of user information are much more valuable than others in terms of their ability to impact ad effectiveness. We also observe that the value of increasing privacy-intrusion exhibits diminishing returns and after the inclusion of temporal variables at the aggregate website level, additional variables do not offer much increase in ad effectiveness.

5. Conclusion

The question of targeting via cookies and privacy violations due to cookie data harvesting has been a challenging policy question. Privacy advocates support strong restriction on cookie data use and advertisers argue such information restriction will limit their ability to innovate. Despite some

papers in this field, no one has assembled a dataset to precise show (i) what information in cookies is relevant for targeting, (ii) and whether targeting based on cookie data is even effective.

Using two very rich and unique field datasets, we fill this research gap. We find that more privacy intrusive information indeed allows for better predictions on user's value but at a decreasing rate. In short, after some cookie data is used, rest of the data does not improve our predictions and restricting their use will not hurt advertiser's ability to target. More specifically, we find that information such as time spent by an individual on different kinds of web pages within the website, does not improve the prediction over and above the accuracy that we get by just using aggregate temporal measures, while they may harm the consumer's privacy perception.

We also find that ads have a small but significant positive effect on purchase probability. But more importantly, we find that targeting is effective. Ads are more effective to users who are more likely to purchase. Thus an advertiser's ability to detect such users generates not only more revenues to the firm, it is also socially beneficial.

We then combine these two results to estimate the value of user information in terms of its effect on sales. We simulate counterfactual privacy policy regimes and calculate the incremental sales that would result from targeted advertising under that regime. We find that only certain kinds of user information increase the ad effectiveness significantly, while others do not have much of an impact.

We believe our research is the first to attempt to quantify the incremental economic value of information that is tracked by cookies and propose a methodology by which policy makers can weigh the costs and benefits of different types of privacy policy regimes.

References

Budak, C., Goel, S., Rao, J. M., & Zervas, G. (2014). Do-Not-Track and the Economics of Third-Party Advertising. *Boston U. School of Management Research Paper*, (2505643).

Goldfarb, Avi, and Catherine E. Tucker. "Privacy regulation and online advertising", *Management Science* 57.1 (2011): 57-71.

Goldfarb, Avi, and Catherine Tucker. "Online display advertising: Targeting and obtrusiveness." *Marketing Science* 30.3 (2011): 389-404.

Hoofnagle, Chris Jay, Jennifer M. Urban, and Su Li. "Privacy and Modern Advertising: Most US Internet Users Want 'Do Not Track' to Stop Collection of Data about their Online Activities." *Amsterdam Privacy Conference.* 2012.

Johnson, Garrett A., Randall A. Lewis, and David H. Reiley. *Add more ads? experimentally measuring incremental purchases due to increased frequency of online display advertising*. Working paper, 2013.

Lambrecht, Anja, and Catherine Tucker. "When does retargeting work? information specificity in online advertising." *Journal of Marketing Research* 50.5 (2013): 561-576.

Lewis, Randall A., Justin M. Rao, and David H. Reiley. "Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising." *Proceedings of the 20th international conference on World Wide Web*. ACM, 2011.

Lewis, Randall A., and Justin M. Rao. "The unfavorable economics of measuring the returns to advertising." *Available at SSRN 2367103* (2014).

Lewis, Randall, and David Reiley. "Retail advertising works! Measuring the effects of advertising on sales via a controlled experiment on Yahoo!." (2009).

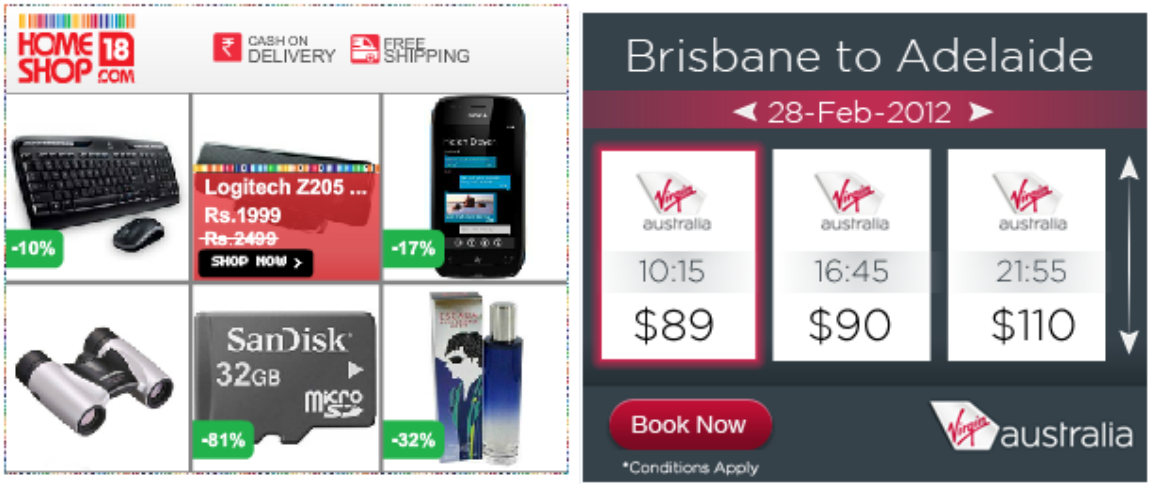
Morales, L. "US internet users ready to limit online tracking for ads." *Gallup Polls* (2010).

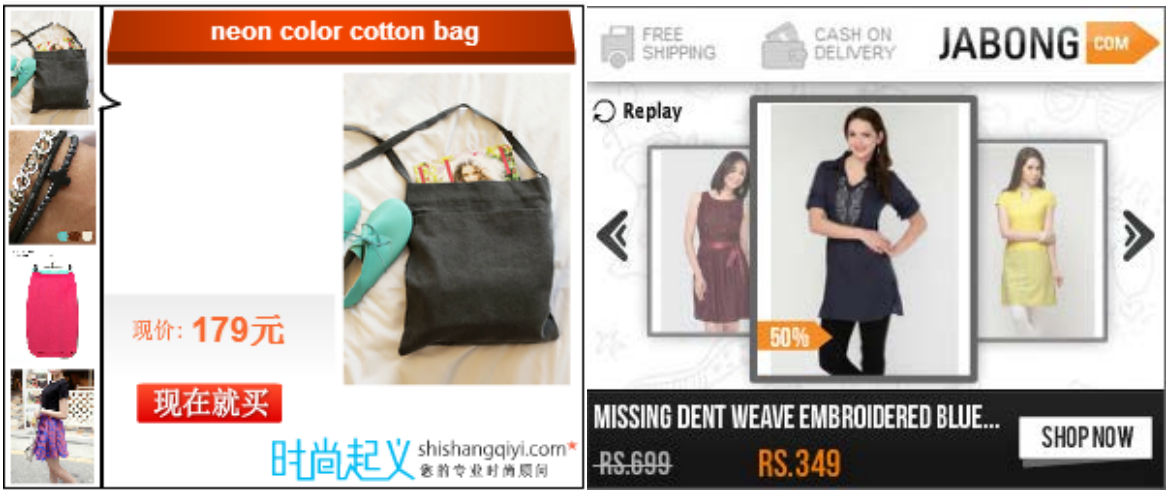
Tucker, Catherine. "Social networks, personalized advertising and privacy controls." *Journal of Marketing Research* (2014).

Turow, Joseph, et al. "Americans reject tailored advertising and three activities that enable it." *Available at SSRN 1478214* (2009).

Appendix I

Personalized targeted ads





Appendix II

In section 4.1, we grouped all the browsing variables into different categories. We then combined these categories to form different models that were used for predicting purchases. In this section we report the AUC results if we had chosen a different grouping scheme.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	Non-behavioral variables	Model 1 + Website level variables	Model 2 + Website Temporal variables	Model 3 + Website product level variables	Model 4 + Product Temporal variables	Model 5 + Last seen/Most seen frequency variables
AUC	0.5654682	0.7529817	0.8128696	0.8134579	0.81296	0.814859

As earlier, we continue to have additional variables added to one model to arrive at the next model. Since we keep adding potentially privacy intrusive variables, each subsequent model is more privacy intrusive than the preceding one.

The difference in this grouping scheme compared to that in section 4.1 is that we now add the website temporal variables (time spent and time since last visit) in Model 3. Thus Model 3 can be interpreted as including only aggregate website level variables, which include the temporal variables. Model 4 adds product level variables which include counts of the number of different category and product pages the individual has visited. Model 5 adds temporal variables that capture the time spent by the individual on different types of pages. Model 6 adds all the other variables that include counts of the last seen and most seen products.

We see that most of the increase in AUC has occurred by Model 3, and subsequent additional variables do not significantly increase the AUC further.