How Effective is Targeted Advertising?

Ayman Farahat
Yahoo!
afarahat@yahoo-inc.com

Michael Bailey
Department of Economics, Stanford University
mcbailey@stanford.edu

ABSTRACT

Advertisers are demanding more accurate estimates of the impact of targeted advertisements, yet no study proposes an appropriate methodology to analyze the effectiveness of a targeted advertising campaign, and there is a dearth of empirical evidence on the effectiveness of targeted advertising as a whole. The targeted population is more likely to convert from advertising so the response lift between the targeted and untargeted group to the advertising is likely an overestimate of the impact of targeted advertising. We propose a difference-in-differences estimator to account for this selection bias by decomposing the impact of targeting into selection bias and treatment effects components. Using several large-scale online advertising campaigns, we test the effectiveness of targeted advertising on brand-related searches and clickthrough rates. We find that the treatment effect on the targeted group is about twice as large for brand-related searches, but naively estimating this effect without taking into account selection bias leads to an overestimation of the lift from targeting on brand-related searches by almost 1,000%.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics
; G.3 [**Probability and Statistics**]: Experimental Design

General Terms

Experimentation, Measurement, Economics

Keywords

Behavioral targeting (BT), Clickthrough rate (CTR), Online advertising, Advertising effectiveness, Field experiments, Selection bias

1. INTRODUCTION

As online advertising is proliferating at an ever increasing pace, clickthrough rates for online advertisements have decreased from 3% to far less than 1% [15]. To improve the effectiveness of their campaigns, advertisers and content providers are increasingly turning to *targeted advertising*, or advertising methods that deliver individually catered advertisements based upon the content of the website, location of the user, browsing history, demographics, the user profile, or any other available information. Purveyors of targeted advertising often promise improved performance, not only in being able to deliver the advertisement to desired user segments, but also increased performance metrics like clickthrough rates (CTR) and sales conversions.

Nevertheless, there are few studies to date that measure the effectiveness of targeted advertising. Given that targeted campaigns carry a premium over other advertising products [5], advertisers are demanding more accurate estimates of the impact of targeting to be able to evaluate whether the additional cost is greater than the marginal return on a targeted advertisement.

Advertisers target their advertisements to the group of users they expect are most likely to respond to the advertising. This provides a major challenge in estimating the effect of targeted advertising because the population is changing simultaneously with the ad, and this *selection bias* will cause any study that naively looks at response lifts between the targeted and untargeted group to greatly overestimate the effect of advertising [4].

To effectively analyze the impact of targeted advertising, we must not only measure the response of the targeted and non-targeted populations to the advertising, but also measure their response in the absence of the advertising intervention, allowing one to measure the *treatment effects* of advertising. There is only value in targeting if the treatment effect on the targeted group is greater than the effect on the untargeted group.

In this paper, we discuss previous theoretical and empirical work on targeted advertising and discuss how our methods account for the selection bias ignored in previous work on targeting. We then introduce a difference-in-differences estimator to evaluate the effectiveness of targeting that controls for the selection bias and using a large-scale natural field experiment involving several online advertising campaigns and a specific interest-based targeting product, we compare our bias-corrected estimates of the impact of targeted advertising with naive estimates. Finally, we estimate a model of targeted CTRs that decomposes the effect of the advertisement, clickiness (the propensity to click on any ad) of users, and brand and category interest of users on targeted CTRs.

We find that brand search lifts from targeting are almost entirely selection bias (77% of the lift on average), but the treatment effect for the targeted population is double that

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/12/04.

of the untargeted population. Clickthrough rate lifts are mostly the treatment effect (only 11% is selection bias), but the median bias-corrected CTR lift is only 1/3 that of the naive CTR lift and we argue that this is a lower bound of the selection bias given the targeting algorithm we analyze. Finally, we find that brand interest is by far the most important determinant of targeted CTRs, greatly outweighing the clickiness of targeted users and the attributes of the advertisement.

2. TARGETED ADVERTISING

Although our evaluation methods could be applied to any type of targeting, we will focus on behavioral targeting (BT) for two reasons: (1) Most of the work in evaluating the effectiveness of targeted advertising has focused on behavioral targeting and (2) since targeted users are chosen based upon similar behavior, traditional measures of advertising effectiveness are very likely to ignore a strong selection bias; the targeted users' behavior is very likely to be highly correlated with the measured response.

Yan, et. al. [23] offer one of the first looks into whether there is any value in targeting in online media. Their goal is to see if targeting ads based upon user behavior leads to a significant improvement in clickthrough rates (CTRs). First, they segment their sample of users into groups defined by similar browsing and query behavior. For each ad in their sample, they find the segment that had the highest CTR on that ad and estimate the potential CTR lift from targeting as:

$$\frac{CTR_{segment} - CTR_{ALL}}{CTR_{ALL}} \qquad \text{(CTR Lift)}$$

where $CTR_{segment}$ is the highest CTR on the ad amongst all segments and CTR_{ALL} is the average CTR on the ad. They find that through segmentation the CTR can be improved by as much as 670% and argue that with more novel segmenting approaches CTRs could be improved as much as 1000%. However, one cannot say whether the increased CTR is because the advertisement was a good fit for that segment or whether it just so happened that the particular segment contained the users with the most clicking and online activity. For example, the user segment with the highest CTR on an ad could conceivably click more on several ads or most ads in a particular category, indicating that segmentation or targeting didn't deliver clicks from users interested in the product promoted by the ad, but just delivered users more likely to click on any ad.

The CTR lift could be a valid measure of the value of targeting if all the advertisers cared about is clicks, but if advertisers care about interest in their category or brand, there is no way to tell how much of this lift was due to a good match between the product in the ad and the interests of the high CTR segment, or whether the segment would have a higher CTR lift for any generic ad. Additionally, because their analysis is done in an ex-post fashion simply selecting the segment with the highest CTR, it ignores the problem that advertisers must choose their targeting model or segment before the ad is shown generously assuming that advertisers always target the optimal segment. A priori, unless most groups of users have similar clicking propensities for all ads, we should expect this methodology to estimate a large CTR lift from targeting for any ad, but this CTR lift overestimates the value of targeting to advertisers, especially if advertisers care about which users are clicking on the ads[1].

Chang and Vijay [7] use historical data on Yahoo! properties to compare the CTR of users who would have qualified for a particular BT category for an ad versus the CTR of all users. For example, a BT category could be users interested in finance, and the authors would then see if those who qualify for that category have a larger CTR on the finance ad then all users. They estimate a variant of CTR Lift:

$$lift_{ad} = \left(\frac{CTR_{ad}^{qualified}}{CTR_{ad}^{all}} - 1\right) \cdot 100 \qquad \text{(CTR Lift 2)}$$

and find that the CTR lift is 39% over typical Yahoo! users and even more on sites with less contextual information like the Front Page (56%) or Mail (61%).

Again, the problem with using this measure of the lift is that users who qualify for the BT category differ systematically from all users and the CTR lift could be just as high on any random advertisement because the users who qualify for the BT segment might have more online activity and are more likely to click on any given advertisement.

This paper extends their work in several ways: (1) We lay down a rigorous theoretical framework and econometric methodology to distill the effectiveness of targeted advertising controlling for selection bias (2) Exploiting a large scale natural experiment that exposes targeted and untargeted users to both targeted and untargeted ads, we can measure how much search and CTR lifts are due to the advertisements, the targeting, and the clickiness of users (3) We provide an empirical model of CTRs to explain the variation in CTRs due to targeting.

3. IDENTIFYING THE IMPACT OF TARGETING USING TREATMENT EFFECTS

A treatment effect (TE) is the average causal effect of some treatment, policy, or program on some measurable outcome of interest, e.g. the effect of a job-training program on future employment rates. Our goal is to measure the treatment effect of targeted advertising keeping in mind that with targeting, the population receiving the treatment might differ from the population not receiving the treatment, and an appropriate methodology must account for this population difference.

Following the standard notation on treatment effects [12], let Y_{i1} be the response of individual i when individual i receives the treatment and Y_{i0} be the response of individual i when individual i is untreated, or assigned to the control group (for example if the outcome of interest is brand related queries then $Y_{i1} = 1$ if the user makes a brand related query after seeing the advertisement and $Y_{i1} = 0$ otherwise).

Let D_i be an indicator variable equal to 1 if individual i

[1] The authors also introduce an F measure to measure the effectiveness of targeting. However, there is no natural interpretation for this ordinal F measure and it cannot inform an advertiser of the marginal revenue from a targeted advertisement. More importantly, using F as our estimate for the value of targeting does not eliminate the selection bias problem; an advertisement might have a very high F measure, but it could be that the same targeted group would have a similar F measure on any ad meaning the users have a high clickiness.

receives the treatment and equal to 0 otherwise (in our case $D_i = 1$ indicates the user saw the advertisement).

We want to measure the impact of the advertisement, or the *individual treatment effect*, $Y_{i1} - Y_{i0}$, for each individual so that we can construct the *average treatment effect* (ATE):

$$\mathbb{E}(Y_{i1} - Y_{i0}) \tag{ATE}$$

(expectation is taken over the population). The ATE is the average differential in response between the users who saw the ad and those who did not. For each individual we observe $Y_i = Y_{i0} + D_i(Y_{i1} - Y_{i0})$, so we can never estimate $Y_{i1} - Y_{i0}$, the individual treatment effect, or the ATE because we can't simultaneously put the individual in the treatment and control. We can't see what the user would do after seeing the ad, and then simultaneously measure what the user would have done had they not seen the ad.

Another widely used measure of the impact of the treatment is the *average treatment effect on the treated (ATET)*:

$$\mathbb{E}(Y_{i1} - Y_{i0}|D_i = 1) = \mathbb{E}(Y_{i1}|D_i = 1) - \mathbb{E}(Y_{i0}|D_i = 1) \tag{ATET}$$

which is the average treatment effect for those who are assigned to receive the treatment. Because $Y_{i0}|Di = 1$ is not observed, the estimation of the ATET is impossible to estimate directly (there is no such thing as a group that is treated and does not receive the treatment).

One approach to measuring the treatment effect is to measure the difference in outcome between the treated and the untreated. The Naive estimator, which compares the average outcome between the treatment and control, can be written as:

$$\begin{aligned} NAIVE &= \mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0) \qquad \text{(Naive)} \\ &= \mathbb{E}(Y_{i1}|D_i = 1) - \mathbb{E}(Y_{i0}|D_i = 0) \\ &= \mathbb{E}(Y_{i1}|D_1) - \mathbb{E}(Y_{i0}|D_1) + \mathbb{E}(Y_{i0}|D_1) - \mathbb{E}(Y_{i0}|D_0) \\ &= \underbrace{\{\mathbb{E}(Y_{i1}|D_1) - \mathbb{E}(Y_{i0}|D_1)\}}_{\text{ATET}} + \underbrace{\{\mathbb{E}(Y_{i0}|D_1) - \mathbb{E}(Y_{i0}|D_0)\}}_{\text{Selection Bias}} \end{aligned}$$

Our Naive estimate is equal to the ATET, plus a term we denote the *Selection Bias*. The selection bias is the difference in response between the selected and unselected populations from being left untreated. The Naive estimator completely ignores what the treated would have done had they not seen the ad, and the untreated had they seen the ad. If there is no difference between the treated and untreated population, for example if D_i was chosen at random through a controlled experiment, then $\mathbb{E}(Y_{i0}|D_1) = \mathbb{E}(Y_{i0}|D_0)$ and the selection bias would be 0. The random assignment of control allows one to just look at difference in outcome between the treated and untreated to measure the treatment effect. In our case, assignment to treatment and control is determined by the targeting criterion of the advertiser which makes D_i far from random.

When the treated and untreated population differ remarkably, the selection bias could be very large. For example, consider site retargeting where the targeted population are those who visited the advertiser's website. These users have already displayed interest in the brand or product being promoted and are much more likely to convert than the general population in the absence of treatment. The Naive estimate ignores the fact that the targeted population is more likely to convert even without seeing the ad and overestimates the ATET by the amount of selection bias.

3.1 Difference-in-differences

The value of targeting to advertisers is the ATE, or how much larger the treatment effect is on the treated than on the untreated. If the treatment effect is the same size for the targeted and untargeted group, then targeting will not improve the marginal revenue of the ad. To get the ATE, we must take the ATET and subtract the ATE on the untreated (ATEU):

$$\begin{aligned} ATE &= ATET - ATEU \\ &= \mathbb{E}(Y_{i1}|D_1) - \mathbb{E}(Y_{i0}|D_1) - ((\mathbb{E}(Y_{i1}|D_0) - \mathbb{E}(Y_{i0}|D_0)) \\ &\equiv (\mu_{11} - \mu_{01}) - (\mu_{10} - \mu_{00}) \qquad \text{(DID)} \end{aligned}$$

Where μ_{kl} is the average response of the group with treatment assignment $D_i = l$ who receive treatment $k = (0, 1) =$ (untreated, treated). To estimate this difference-in-differences (DID), we must see how targeted and untargeted users respond to the advertising, and how targeted and untargeted users respond in the absence of advertising. We can then find the effect of the treatment on the targeted group, and difference out the treatment effect on the untargeted group to find the marginal impact of showing an advertisement to the targeted group over the untargeted group.

We can rearrange the DID and write the ATE as:

$$ATE = (\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00}) \tag{DID2}$$

(DID) finds the ATE by finding the ATET and then subtracting off the ATEU. (DID2) offers an alternative interpretation of (DID), we can obtain the DID estimate by first estimating the difference in response between the targeted and untargeted population to the treatment. However this difference might be due solely to differences in the populations, and have nothing to do with the advertising, so we must correct for this bias by differencing $(\mu_{01} - \mu_{00})$, the difference in response between the targeted and untargeted group in the absence of the advertising intervention, which tells us how much of the difference is due solely to the differences between the population. It should be noted that matching estimators, which match similar users who received different treatments, will only help in estimating the ATET; to obtain the ATE we need a proper difference-in-differences to account for the selection bias.

The DID estimator also falls naturally out of an econometric model as follows.

4. ECONOMETRIC MODEL

We model the process of users responding to an advertisement as a repeated Bernoulli trial with probability of success μ. This probability depends on whether the user saw the ad or not, and other observable characteristics of the user, so we assume that users who are observably identical have the same probability of conversion.

Suppose we run an experiment and randomly expose some subset of the population to the advertisement, and leave some part of the population unexposed. Additionally, there is a subset of the population that the advertiser is targeting for the advertising campaign, but exposure to the ad is random and irrespective of targeting. Let $y_i = 1$ if user i converts and $y_i = 0$ otherwise. If user i has probability of conversion, μ, then :

$$\begin{aligned}\Pr(y_i = 1) &= \mathbb{E}(y_i) = \mu \\ Var(y_i) &= \mu(1-\mu)\end{aligned}$$

If Y is the number of users who convert out of n total users (who convert with probability μ), a sum of the Bernoulli successes, then Y is distributed binomial with mean $n\mu$ and variance $n\mu(1-\mu)$. Therefore, the expected conversion rate, $\mathbb{E}(\frac{Y}{n})$, has mean μ and variance $\frac{\mu(1-\mu)}{n}$.

We model the expected value, $\mathbb{E}(\frac{Y}{n}) = \mu$, as a linear combination of the independent variables, $\mathbf{X}$ (which includes whether the user is targeted and whether the user is exposed to the ad) through a link function $f()$:

$$\mathbb{E}\left(\frac{Y}{n}\right) = \mu = f^{-1}(\mathbf{X}\beta)$$

Since $\frac{Y}{n}$ is distributed binomial, the appropriate link function is the logit, $f(x) = \frac{x}{1-x}$. If the number of users in the experiment is large, $\frac{Y}{n}$ is approximately distributed normal with mean μ and variance $\frac{\mu(1-\mu)}{n}$. The normal distribution corresponds to the link function $f(x) = x$. Additionally, if the probability of conversion is small, $\frac{Y}{n}$ is also approximately distributed Poisson with mean μ and variance $\frac{\mu}{n}$ (if μ is small then $\mu \approx \mu(1-\mu)$). The Poisson link function is $f(x) = \ln(x)$.

$$\ln\left(\frac{\mu}{1-\mu}\right) = \mathbf{X}\beta \qquad \text{(Binomial)}$$

$$\mu = \mathbf{X}\beta \qquad \text{(Normal)}$$

$$\ln(\mu) = \mathbf{X}\beta \qquad \text{(Poisson)}$$

Ignoring other independent variables, suppose the probability the user clicks on the advertisement is expanded as:

$$\begin{aligned}f(\mu_i) &= \mathbf{X}\beta \\ &= \beta_0 + \beta_1 Ad_i + \beta_2 Target_i + \beta_3 Ad_i \cdot Target_i\end{aligned}$$

Where Ad_i is a dummy variable equal to unity if the user i saw the ad and $Target_i$ is a dummy variable that is equal to unity if user i is targeted for the ad, and they are both equal to 0 otherwise.

β_0 is the baseline conversion percentage, or the conversion rate of users who don't see the ad and are untargeted.

β_1 is the difference in the probability of conversion between those seeing the ad and those not seeing the ad, it is the marginal effect of the ad holding targeting constant.

β_2 is the difference in probability of conversion between the targeted and untargeted group. This is the Selection Bias.

β_3 is the interaction term between being shown the targeted ad as well as being in the targeted segment. This measure is the marginal increase in the probability of conversion when the user is shown the ad and is part of the targeted group or the (ATE). This is the value of targeting to advertisers answering, "how much larger is the treatment effect on the targeted group?"

Note that $\beta_1 + \beta_3$ is the ATET and β_1 is the ATEU and that when computing averages over populations we would find that:

$$\begin{aligned}f(\mathbb{E}(y|Ad = 1, Target = 1)) &= \beta_0 + \beta_1 + \beta_2 + \beta_3 \\ f(\mathbb{E}(y|Ad = 1, Target = 0)) &= \beta_0 + \beta_1 \\ f(\mathbb{E}(y|Ad = 0, Target = 1)) &= \beta_0 + \beta_2 \\ f(\mathbb{E}(y|Ad = 0, Target = 0)) &= \beta_0\end{aligned}$$

If we were to use the naive estimates of the impact of targeting we would find that :

$$\begin{aligned}NAIVE &= f(\mathbb{E}(y|Ad = 1, Target = 1)) \\ &\quad - f(\mathbb{E}(y|Ad = 0, Target = 0)) \\ &= \beta_0 + \beta_1 + \beta_2 + \beta_3 - (\beta_0) \\ &= \beta_1 + \beta_2 + \beta_3 \\ &= (\beta_1 + \beta_3) + \beta_2 \\ &= \text{Treatment Effect} + \text{Selection Bias}\end{aligned}$$

If we had shown all targeted users the advertisement, and withheld the ad from all the untargeted users, we would have a fundamental identification problem; whenever $Target = 1$, we would have that $Ad = 1$, so there is no way to separately identify β_2 and β_3. By determining who sees the ad at random however, we can properly identify β_3 using (DID):

$$\begin{aligned}DID &= f(\mathbb{E}(y|Ad = 1, Target = 1)) \\ &\quad - f(\mathbb{E}(y|Ad = 0, Target = 1)) \\ &\quad - f(\mathbb{E}(y|Ad = 1, Target = 0)) \\ &\quad + f(\mathbb{E}(y|Ad = 0, Target = 0)) \\ &= \beta_0 + \beta_1 + \beta_2 + \beta_3 - (\beta_0 + \beta_2) - (\beta_0 + \beta_1) + \beta_0 \\ &= \beta_3\end{aligned}$$

which yields an estimate of the desired parameter.

The test statistic for the difference in differences estimate of the impact of targeting for the three link functions is:

$$\begin{aligned}&\left[\ln\left(\frac{\mu_{11}}{1-\mu_{11}}\right) - \ln\left(\frac{\mu_{10}}{1-\mu_{10}}\right)\right] \\ &- \left[\ln\left(\frac{\mu_{01}}{1-\mu_{01}}\right) - \ln\left(\frac{\mu_{00}}{1-\mu_{00}}\right)\right] \\ &= \ln\left(\frac{\left(\frac{\mu_{11}}{1-\mu_{11}}\right)}{\left(\frac{\mu_{10}}{1-\mu_{10}}\right)} \div \frac{\left(\frac{\mu_{01}}{1-\mu_{01}}\right)}{\left(\frac{\mu_{00}}{1-\mu_{00}}\right)}\right) = \beta_3 \qquad \text{(Logit)}\end{aligned}$$

$$(\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00}) = \beta_3 \qquad \text{(Identity)}$$

$$(\ln\mu_{11} - \ln\mu_{10}) - (\ln\mu_{01} - \ln\mu_{00}) = \ln\left(\frac{\frac{\mu_{11}}{\mu_{10}}}{\frac{\mu_{01}}{\mu_{00}}}\right) = \beta_3 \quad \text{(Log)}$$

When using the identity link function, we can interpret β_3 as the marginal increase in conversion probability. The interpretation of β_3 is not as straightforward for the other link functions as the marginal increase is in terms of the log odds ratio. As an alternative to the differences in differences estimator, we use the following approximation to the logarithmic function, $\ln(1 + x) \approx x$ for x small, to show that a quotient-in-quotients (QQ), which has a more natural interpretation than the other estimators for β_3, can sometimes approximate the Poisson link estimator:

$$\ln\left(\frac{\frac{\mu_{11}}{\mu_{10}}}{\frac{\mu_{01}}{\mu_{00}}}\right) \approx \left(\frac{\frac{\mu_{11}}{\mu_{10}}}{\frac{\mu_{01}}{\mu_{00}}}\right) - 1 \text{ for } \left(\frac{\frac{\mu_{11}}{\mu_{10}}}{\frac{\mu_{01}}{\mu_{00}}} - 1\right) \text{ small} \qquad \text{(Quotient-of-quotients)}$$

The (Quotient-of-quotients) is interpreted as percentage deviations from the base conversion rate. The naive quotient, $\frac{\mu_{11}}{\mu_{10}}$, is the ATET in percentage terms, and when divided by $\frac{\mu_{01}}{\mu_{00}}$ yields the ATE in percentage terms.

The main takeaway from the model is that to properly evaluate the effectiveness of targeted advertising, the appropriate experiment would be to show the advertisement to some members of the untargeted sample, and to withhold the advertisement from some members of the targeted sample, to see how they respond in the presence/absence of advertising. Only then can the proper difference-in-differences be estimated. The samples do not have to be evenly sized, but the power of the test to determine significant differences is dependent on the sample sizes of the different groups.

5. NATURAL FIELD EXPERIMENT

We exploit a natural field experiment from the large rectangular ad unit on the Yahoo! Front Page (www.yahoo.com). Yahoo! Front Page advertisements are sold in *roadblocks* (every user is delivered the advertisement) and occasionally are split evenly between two advertisers such that a visitor will be shown an advertisement from the first advertiser if the time of arrival to the Front Page is on an even second, and from the second advertiser if the arrival time is on an odd second, regardless of the characteristics of the visitor. An impression for one of the two advertisers is shown for every visit to the front page on that particular date. This is known as a *front page split*.

This provides the perfect experimental setup to compute the Naive and DID estimates of the advertising campaigns. For each front page split we conduct two experiments; we pick one of the two campaigns as our target campaign, and define the other campaign in the split as our control campaign. We then switch the roles of the advertisers, giving us two experiments per front page split. We observe how the targeted and untargeted populations respond to the test advertisement and the control advertisement to compute the DID and Naive estimates of targeting.

Although none of the front page campaigns we analyze are a targeting campaign, for each advertiser we chose the interest category they would have been most likely to target. For each user we observe their behavioral targeting profile and know to what behavioral targeting segments they belong, so we can compute the hypothetical search lift that would have occurred had the advertisement run as a targeting only campaign for the advertiser's interest category. For example, if the advertisement is finance related, we can identify which users belong to the finance BT segment and then compute the naive and DID estimators of the search lift for that category of users.

Motivated by recent results that highlight the impact of search on advertising metrics [17], we begin by examining the impact of display ads on both brand and generic category search terms. We follow with an analysis of clickthrough rate lifts for the same campaigns. Analyzing search lifts follows the setup of the econometric model exactly. We assume that each user has a constant probability of making a brand (or category-related) search, and this probability differs upon whether the user sees the advertisement and is in the target advertisers BT category.

Our results are specific to the targeting product chosen and we would expect different results for different targeting products and different kinds of targeting. Different ways of segmenting and clustering users might induce larger amounts of selection bias when computing the effectiveness of targeting. For our experiment, we chose to use an interest-based targeting product, Yahoo!'s BT Engager. BT Engager places users into broad interest-based categories (like *Chinese language* or *finance*) on the basis of searches, pageviews, clicks, and other browsing behavior. Because this model is meant to identify broad interest instead of maximizing click-through rate, it provides a conservative estimate of the selection bias that one would find using more aggressive CTR maximizing targeting strategies.

5.1 Targeting's Impact on Search

All data, tables, and results from the paper can be found in the online appendix included in the supplementary materials for the paper. The experiment includes 18 advertising campaigns on 9 front page splits from May through August, 2011 (See table 1)[2].

For each user, we track searches made on Yahoo! search by the user after their first visit to the front page, but before their second visit (the second impression) and we also exclude searches made 10 minutes after they view the advertisement. The choice of 10 minutes was motivated by the fact that most searches occur within 10 minutes of the first visit to the Yahoo! home page. Searches after 10 minutes only add noise to the estimates. Removing searches made after the second impression allows us to cleanly identify which advertisement influenced the search and to ignore frequency effects [3]. For the 18 campaigns there were 332 million such impressions with an average of about 18.4 million impressions per campaign.

We define a target segment search as a search that is related to the category of the target advertisement[4]. Similarly, a control segment search is a search related to the category of the control advertisement. For the brand related searches, we identified the most salient brand associated with each advertisement and define a brand search (either target or control) as a search that includes the brand name.

We find that the ads have a sizeable effect on search; 9/18 ads have a statistically significant and positive effect on brand searches, and only one of the ads had a negative effect

[2] The random assignment of ads (even second arrivals to advertisement 1, odd second arrivals to advertisement 2), failed for one hour during the 06/20 front page split which is why the impressions for the two advertisers is not evenly distributed. Because we have no reason to assume that users arriving during that one hour differ substantially from the rest of the users, it shouldn't influence the outcome of the two experiments for that day.

[3] we also repeated the experiment including all searches and the results are substantively the same

[4] Precisely, Yahoo! has compiled a list of canonical searches for every BT category, and we define a segment search as a search that is in the top 100 search terms for the BT category of the target advertiser. For example, if auto insurance was our target advertisement campaign with a BT category of 'insurance', the top 100 canonical search terms would include terms such as 'insurance', 'geico', 'progressive', 'auto insurance', and 'prudential'.

Table 1: Targeted Brand Search Lift and CTR Lift, Diff-in-diff Lifts, and % of Naive Lift that is Selection Bias for the 18 Campaigns in Analysis.

	Brand Search Lifts			CTR Lifts		
Target Category	**Naive Lift**	**Diff-in-diff**	**SB % of Lift**	**Naive Lift**	**Diff-in-diff**	**SB % of Lift**
Credit Card	1,647%***[a]	0.00%	**101%**	159%***	148%***	**7%**
Insurance 1[b]	773%***	0.02%	**83%**	123,375%***	119,901%***	**3%**
Credit Card	218%***	0.01%	**90%**	777,474%***	761,598%***	**2%**
Insurance 2	113%***	-0.01%	**174%**	22%***	22%***	**-2%**
College	477%***	0.00%	**87%**	17%**	4%	**77%**
Insurance 1	738%***	0.01%	**92%**	39%***	8%***	**81%**
Notebooks	679%***	0.03%	**93%**	136%***	136%***	**1%**
Digestive System	-100%**	-0.01%	**-**	126%*	-44%**	**135%**
Notebooks	2,010%***	0.11%***	**77%**	5,650%***	5,643%***	**0%**
Insurance 1	618%***	0.01%	**92%**	31%***	37%	**-18%**
Reality TV	-80%**	-0.01%	**-18%**	3%	3%	**22%**
Credit Card	-40%**	-0.01%	**41%**	79%***	276%***	**-250%**
Notebooks	1,148%***	0.09%***	**42%**	207%***	-46%***	**122%**
Adventure Movies	1,281%***	0.38%***	**62%**	39%***	38%***	**1%**
Adventure Movies	1,589%***	0.46%***	**1%**	183%***	171%***	**7%**
Insurance 1	730%***	-0.01%	**109%**	131,134%***	131,385%***	**0%**
College	510%***	0.00%	**78%**	14%**	22%***	**-55%**
Insurance 1	663%***	0.00%	**102%**	52%***	21%***	**59%**
MEAN	721%	0.06%	**77%**	57,708%	56,629%	**11%**
MEDIAN	672%	0.00%	**87%**	102%	37%	**2%**

The Naive Estimate is the CTR difference between the target BT segment and all users not in the target BT segment on the target ad (both excluding the control BT segment).

The Diff-in-diff is the difference between the ad impact on the targeted group, and the ad impact on the untargeted group

SB % of Lift is the percent of the naive lift that can be attributed to selection bias

[a] Two sided t-test p-value using normal approximation; * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; the null hypothesis is that the estimate equals 0.

[b] One advertiser, designated "Insurance 1", appears 5 times in our sample with similar advertisements. All other brands are unique.

on searches for the brand of the ad. The average population treatment effect lift in brand related searches is 44.7%. For those in the targeted BT category, four of the ads have a positive and statistically significant effect on brand searches and the average treatment effect search lift is 1.8% and 79.9% for segment and brand searches respectively.

Table 2 presents the average search lifts for category and brand searches for the targeted and untargeted ads and for the targeted and untargeted populations. The Naive estimates are the search lifts between the target group seeing the target ad and the untargeted group seeing the target/control ad (columns of table 2). The naive estimate of the effect of targeting on searches is a lift of almost 3,000% for category searches and 800% for brand searches! An overestimate on the order of 1000%. The targeted group is much more likely to make a brand related search no matter what ad is seen, so the lift between the targeted and untargeted group is mostly selection bias (median of 87% selection bias). Table 1 shows the naive brand search lift and DID brand search lift for each of the 18 campaigns.

The average DID estimate of the effectiveness of targeting is -0.013% for a category search and 0.067% for a brand search, or median lifts of about 4% and 51% respectively. The selection bias accounts for 98% of the naive search lift for segment searches, and 77% of the search lift for brand searches. Three of the campaigns are associated with a negative naive search lift, i.e. the targeted group made fewer searches after seeing the ad. This gives evidence that interest category targeting does not always yield the optimal target audience, although since 2 of the 3 campaigns occurred on 06/23 there could be an anomaly on that date.

There is an enormous variation in the naive lifts, and to a lesser extent the DID lifts between the campaigns. One reason for this variation is that we throw out all users who belong to the BT group of the control advertiser. Because there are different levels of overlap between the target and control BT group, we keep a majority of the target group for some campaigns and not others. Another reason for the variation is that there is variation in the BT engager models. For segments in which demand is high, there is likely more sophistication in the modeling and parameters are adjusted so that supply meets demand.

5.2 Analyzing Clickthrough Rates with Treatment Effects

The treatment effects results we derived earlier are not as portable to a clickthrough rate analysis. Users can make a search or make some kind of conversion without seeing the advertisement, but they cannot click on an advertisement without seeing the advertisement. The Naive estimator can no longer be used because we cannot observe clicks from the untreated group.

In this analysis, we define the naive estimator as the difference in CTR between the targeted group and the untargeted group on the targeted advertisement. This is the CTR Lift estimate discussed previously ([23] and [7]).

Suppose we also showed the targeted and non-targeted population a placebo, or control ad. An ad for which targeting should not be a factor in generating a CTR lift between the targeted and non-targeted population. If there is any CTR lift between the groups on this placebo ad, then it can be attributed to a higher click propensity by the targeted population, and not due to the ad being a better 'fit' or being more appropriate for the targeted group. This is how we define selection bias in this problem, the CTR lift between the targeted and non-targeted group on a placebo ad for which the lift is solely attributed to a higher click propensity (perhaps because of more happy clickers in the targeted population).

The ideal experiment then involves showing target ads and placebo ads to both the targeted and non-targeted populations. The value of targeting is the DID estimator:

$$\begin{aligned} DID = & f(\mathbb{E}(y|Ad = 1, Target = 1)) \\ & - f(\mathbb{E}(y|Ad = 1, Target = 0)) \\ & - f(\mathbb{E}(y|Ad = 0, Target = 1)) \\ & + f(\mathbb{E}(y|Ad = 0, Target = 0)) \\ & = NAIVE - SELECTION\ BIAS \end{aligned}$$

The value of targeting is the CTR lift subtracting the baseline click propensity of the targeted group. Another way to get this DID is to first compute the CTR lift of the targeted group between the targeted and untargeted ad. However, this difference could be attributed solely to ad differences, so we must difference form this the CTR lift of the untargeted group between the targeted and untargeted ad as a measure of the baseline ad difference.

There are some important caveats/assumptions with this approach:

(1) We cannot estimate the ATET or any ad impact measures since it is impossible to click in the absence of seeing the ad; the impact of the ad on clicks has no meaning.

(2) When the DID estimator is 0 for the search/conversion lift, it means that the targeted group receives no extra lift from the advertising above and beyond the lift in the general population, this tells us that employing targeting is not any more effective in generating searches/conversions than a run-of-network campaign. In the CTR case, if the DID estimator is 0, it has the different interpretation that the CTR lift on the target ad is identical to the CTR lift on the placebo ad, but depending on the advertisers goals this does not necessarily mean that targeting is not effective.

(3) This strategy relies on the placebo ad being an ad for which any CTR difference is due to population differences in propensity to click only. This assumption fails if preferences for the target and control advertisement differ by treatment group. As an example, suppose that the target ad is for cologne and the placebo ad is for perfume, and the targeted segment is 20-30 year old males. The targeted group is less likely to click on the placebo advertisement, but this is not because of differing click propensities, it is because of differing tastes, which is the problem we are trying to solve on the target ad, how to separate tastes from *clickiness*.

Formally, this assumption states that (using our regression methodology) there are no interaction terms of the form $\beta_4(1 - Ad_i) \cdot (1 - Target_i)$, such that the targeted users respond differently to the placebo ad compared to the general population for reasons not related to clickiness. Similarly, if the placebo ad is very similar to the target advertisement, the targeted group might be more likely to click on it than the general population due to tastes and not due to clickiness.

The front page split is an ideal venue for testing the naive and DID estimators for CTR lifts. Because the ads are assigned at random, we can define one of the two ads as the

Table 2: Average Search Lifts Between Populations and Advertisements.

	Category Searches			Brand Searches		
	Target Ad	**Control Ad**	**Lift**	**Target Ad**	**Control Ad**	**Lift**
Targeted	0.86%	0.87%	**1.8%**	0.19%	0.13%	**79.9%**
Untargeted	0.03%	0.03%	**2.7%**	0.02%	0.02%	**44.7%**
Lift	**3,157%**	**3,272%**		**896%**	**742%**	

Lifts under the columns are the search lifts between the targeted group and untargeted group for the ad in the column. Lifts after the rows are the search lifts between the targeted and untargeted ad for the given group (targeted/untargeted).

target ad and the other as the placebo ad and estimate the naive and DID estimator for CTR lift.

5.3 CTR Results

We restrict our attention to only the first advertisement served so we can measure the marginal impact of a singular impression on clicks[5].

The 18 campaigns had about 18.4 million unique users with a mean CTR of 0.15%. Table 1 shows the naive CTR lift and DID CTR lift for the 18 campaigns. The median naive CTR lift between the targeted and non-targeted group is 102%, and the DID estimator is 89% of the Naive estimate, meaning that selection bias only accounts for 11% of this CTR lift on average. For the ads in our analysis, the naive CTR lift is a good approximation of the treatment effect lift. The Quotient-of-quotients model yields similar results.

There are a few reasons our estimates of the selection bias is small for CTR estimation. It could be that the BT engager model is less prone to assigning users with high *clickiness* to lots of BT categories. 99.5% of our sample of users is included in at least one BT category and the BT categories only overlap by 5% on average, thus a very small portion of our sample is users belonging to several BT categories generating lots of clicks and driving up the selection bias. One of the reasons the search selection bias is high is that users are assigned to BT categories based on both predicted and observed category search patterns, but this phenomenon does not manifest itself for CTR because clicks are a much less common occurrence and the model is only based loosely on CTR.

Another explanation could be that for several of the ads in our sample there was a negative correlation in tastes between the two ads in the front page split such that the target BT's *clickiness* was outweighed by their distaste for the placebo ad. This would be supported by the fact that there is such a small overlap between BT category assignment. However, the categories appear to be unlikely to generate such a negative correlation in tastes.

6. A MODEL FOR CLICKTHROUGH RATES

After accounting for the clickiness of the targeted group, we still find large CTR lifts on front page advertisements. Can we attribute the residual lift to interest in the brand or category? To make this causal claim we need to lay down a behavioral model of clicking that describes why the targeted group is more prone to click on an advertisement than the general population of users. We posit that CTRs for a targeted group of users is dependent on three things:

(1) Creative: Attributes of the ad itself that drives a higher CTR for everyone, like quality of the advertisement or general appeal of the brand. This can be measured by the CTR of all users on the ad.

(2) Clickiness: Click propensity of the group, this captures how much more likely the targeted group clicks on any ad. We measure clickiness by the selection bias, or how much more likely the target group clicks on the placebo ad.

(3) Interest: How much additional interest does the targeted group have for the brand or category. We measure this by looking at the lift in category or brand searches for the target group after seeing a placebo ad which is a measure of the pre-determined interest.

We follow the steps outlined in our econometric setup and estimate this relationship as a generalized linear model. The CTRs are modelled as being distributed Gaussian, Binomial, or Poisson by choosing the appropriate link function with linear condition mean:

$$\begin{aligned} f(\mu_i) &= \mathbf{X}\beta \\ &= \beta_0 + \beta_1 Creative_i + \beta_2 Clickiness_i \\ &\quad + \beta_3 CategoryInterest_i + \beta_4 B\,\mathrm{rand}\,Interest_i \end{aligned}$$

The appropriate link function (or canonical link function) for the Gaussian, Binomial, and Poisson distributions are the identity function, logit function, and log function respectively. We also run a Box-Cox regression [Box & Cox, 1964] to determine the best power transformation of the dependent variable that would satisfy the linear conditional mean assumption. The regression suggests the data is best modelled as log-linear, with an estimated lambda close to 0 ($\lambda = -0.1$), i.e. Poisson. We normalize all independent variables to have mean 0 and a standard deviation of 1 (the independent variables are standard deviations from the mean) and estimate all models with robust standard errors.

For the Poisson and Binomial models, we present exponentiated coefficients. Coefficients for the poisson model can be interpreted as the percentage change in the CTRs per standard deviation increase in the independent variable minus one. For the Binomial model, the interpretation is the percentage change in the odds $\left(\frac{CTR}{1-CTR}\right)$, not percentage change in CTRs; however, since $CTR \approx 0$, the odds is approximately equal to the probability, $\frac{CTR}{1-CTR} \approx CTR$, so the same interpretation can be applied. The coefficients are not exponentiated for the Gaussian model and are the standard marginal effects.

The results of the regression are presented in table 3.

[5] the results from looking at all front page impressions are substantially identical and will be provided upon request.

Brand interest by far is the most important determinant of segment CTR and is statistically significant in two of the models (in all the models using one-tail p-values). A one standard deviation increase in brand interest leads to a 138% increase in CTRs in the Poisson model.

Table 3: Regression Results for CTR Model.

	Binomial	Poisson	Gaussian
Constant			0.006*** (0.001)
Creative	0.69 (0.2)	0.69 (0.2)	-0.001 (0.003)
Clickiness	1.15 (0.23)	1.14 (0.23)	0.0003 (0.001)
Category Interest	1.21 (0.44)	1.21 (0.44)	0.0007 (0.002)
Brand Interest	2.4* (0.88)	2.38* (0.86)	0.006 (0.001)
Link	Logit	Log	Identity
N	18	18	18

Exponentiated coefficients reported for Binomial and Poisson models.
Two sided t-test p-value; * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; the null hypothesis is that the estimate equals 0.
Robust standard errors shown in parentheses.

Clickiness and Category Interest are both positive, but not statistically different from 0. Creative is negative but not statistically significant. Although this seems to suggest that better ads lead to lower CTRs for the targeted group (once controlling for interest and clickiness), it is probably more likely the case that the Creative variable is better interpreted as a measure of general interest in the advertisement, and the more broad appeal an ad has, the lower precision there is in the targeted category. Controlling for brand and category interest, targeting will not be as precise for a general appeal ad like a blockbuster movie.

The target population responds to the ad similarly to the general population once we've accounted for clickiness and category interest, but without accounting for brand interest which drives most of the variation in CTRs. Our sample of creatives is small (n=18), limiting the general applicability of our results, but it presents strong evidence nonetheless that brand interest is a significant determinant of variation in targeted CTRs.

7. MARGINAL VALUE OF A TARGETED ADVERTISEMENT

If the cost per 1000 impressions (CPM) is $1, for the campaigns we analyze it would cost an advertiser on average 72 cents per click and $5.12 for each brand related search. If targeting was used, and CPM for targeted impressions was also $1, a click would only cost 16 cents and a brand related search would only cost 53 cents.

However, for searches, the figure advertisers care about is not the cost per search, but the marginal cost of a brand search, or how much it would cost the advertiser to induce someone to search for the brand by displaying an ad. The marginal cost of a brand-related search on Yahoo! search from any user is $15.65, but is only $1.69 for a targeted user. This only considers searches on Yahoo! and overstates the cost of all brand-related searches on any search engine. Unless showing targeted advertising is 9 times more expensive, targeted advertising is more cost effective in generating brand searches.

If all an advertiser cares about is clicks, and not about the clickiness of the targeted group, then targeted advertising is more cost effective at generating clicks as long as it is no more than 4.5 times as expensive as displaying the ad to everyone. Given an industry average of a 3 times price premium for targeting [5], we might conclude targeting is more cost effective, but of course this depends greatly on the targeting product.

If advertisers only value clicks that aren't due to selection bias, for example if they have a niche product and they only want clicks from those who are uniquely interested in their category, then it would cost the advertiser 22 cents per click, or about 1/3 the cost of a click for a run-of-network campaign.

8. CONCLUSION AND IMPLICATIONS FOR ADVERTISERS

We conclude by discussing what advertisers can learn from our model and results. When computing the effectiveness of a targeted advertising campaign, it is critical to not only compare how the targeted and non-targeted populations respond to advertising, but how they respond in the absence of advertising. This is because the targeted segment is more likely to convert in the absence of advertising than the untargeted segment, and to truly measure the effect of advertising this selection bias must be accounted for.

We find large selection bias in brand related search lifts. If a comparison was made between the searches of the targeted and untargeted group after seeing an advertisement, we would conclude the advertisement lifted brand searches 721% on average. In reality, that same lift is observed after seeing an unrelated advertisement, the true effect of advertising once we account for this lift is around 79%.

We find that for an interest-based behavior targeting model, selection bias or clickiness of the targeted group can only account for about 11% of the CTR lift; we believe this is more a function of a targeting product that is less prone to selection bias and the results are a conservative estimate for a more aggressive behavioral targeting model that seeks to maximize CTRs. If advertisers are discriminate about the types of users clicking on their ads, and not just the number of clicks, our methodology offers insight into how much of the CTR lift from the targeting group is unique to their advertisement category.

When we model CTRs for the targeted groups as a function of advertisement characteristics, clickiness of the group, and pre-determined interest in the category and brand of the advertisement, we find that brand interest overwhelming explains variation in CTRs, so the lift in CTRs from targeted users is being driven by identifying users who have interest in the brand, and not from characteristics of the advertisements. Once we control for category interest and clickiness, targeted users respond to the ad like the general population.

This study opens up several questions about the effectiveness of targeted advertising. Advertisers are seeking more and more to target their ads to the segments most likely to convert as a result of the advertising; however, this strategy may not be cost effective as this segment is likely to convert in the absence of any advertising. Our results indicate that more sophisticated targeting algorithms might not gain, and might even harm, the advertiser as those seeing the ad would convert in the absence of advertising. Targeted advertising is projected to grow at an enormous rate; we hope research on targeted advertising keeps up the pace.

9. REFERENCES

[1] B. Anand and R. Shachar. Targeted advertising as a signal. *Quantitative Marketing and Economics*, 7(3):237–266, 2009.

[2] J. Angrist and G. Imbens. Two-Stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995.

[3] J. Angrist, G. Imbens, and D. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, June 1996.

[4] K. Bagwell. Chapter 28 the economic analysis of advertising. *Handbook of Industrial Organization*, 3:1701–1844, 2007.

[5] H. Beales. The value of behavioral targeting. *Network Advertising Initiative*, 2010.

[6] G. Box and D. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, page 211–252, 1964.

[7] K. L. Chang and V. K. Narayanan. Performance analysis of behavioral targeting at yahoo! Technical report, Advertising Sciences, Yahoo! Labs, 2010.

[8] J. Chen and J. Stallaert. An economic analysis of online advertising using behavioral targeting. Technical report, University of Connecticut, 2010.

[9] T. Chen, J. Yan, G. Xue, and Z. Chen. Transfer learning for behavioral targeting. In *Proceedings of the 19th international conference on World wide web*, pages 1077–1078, 2010.

[10] Y. Chen, D. Pavlov, and J. F. Canny. Large-scale behavioral targeting. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 209–218, 2009.

[11] E. Gal-Or, M. Gal-Or, J. May, and W. Spangler. Targeted advertising strategies on television. *Management Science*, 52(5):713–725, May 2006.

[12] G. Imbens and J. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

[13] G. Iyer, D. Soberman, and M. V. Boas. The targeting of advertising. *Marketing Science*, 24(3), 2005.

[14] M. Joo, K. Wilbur, and Y. Zhu. Television advertising and online search. *Available at SSRN: http://papers. ssrn. com/sol3/papers. cfm*, 2010.

[15] P. Kazienko and M. Adamski. AdROSA-Adaptive personalization of web advertising. *Information Sciences*, 177(11):2269–2295, June 2007.

[16] A. Lambrecht and C. Tucker. When does retargeting work? timing information specificity. *SSRN eLibrary*, July 2011.

[17] J. Lecinski. Zero Moment of Truth. `http://www.zeromomentoftruth.com/`

[18] R. A. Lewis, J. M. Rao, and D. H. Reiley. Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th international conference on World wide web*, page 157–166, 2011.

[19] P. Melville, S. Rosset, and R. D. Lawrence. Customer targeting models using actively-selected web content. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 946–953, 2008.

[20] D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[21] D. Vakratsas, F. Feinberg, F. Bass, and G. Kalyanaram. The shape of advertising response functions revisited: A model of dynamic probabilistic thresholds. *Marketing Science*, 23(1):109–119, Jan. 2004.

[22] T. Wiesel, K. Pauwels, and J. Arts. Practice prize Paper-Marketing's profit impact: Quantifying online and off-line funnel progression. *Marketing Science*, 30(4):604–611, 2011.

[23] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web*, pages 261–270, 2009.